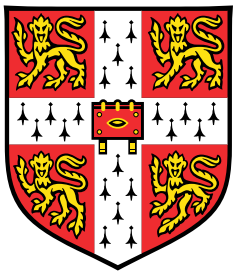
Mapping the Genomic Context of Mutagenesis

A multidimensional machine learning approach.



Harald Sager Vöhringer

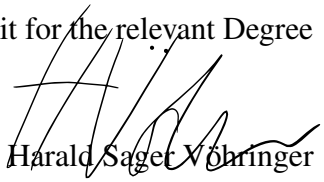
European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Christ's College September 2020

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.



Harald Sager Vöhringer
September 2020

Mapping the Genomic Context of Mutagenesis

Harald Sager Vöhringer

The accumulation of genomic mutations leads to the formation of cancer. For this reason, many efforts have been undertaken to characterise mutational processes in terms of their genomic imprints. A particularly successful approach is matrix-based mutational signature analysis, which identifies prototypical mutation patterns by applying non-negative matrix factorisation to catalogues of single nucleotide variants and other mutation types. However, mutagenesis is a multifaceted event that is affected by the genomic organisation of DNA and cellular processes such as transcription, replication, and DNA repair processes. Moreover, since many mutational processes also generate characteristic multi nucleotide variants, insertion and deletions, and structural variants, it appears valuable to jointly deconvolve broader mutational catalogues to better understand the complex nature of mutagenesis.

In this thesis, I present TensorSignatures, an algorithm to learn mutational signatures jointly across different variant categories as well as their genomic localisation and properties. The analysis of 2,778 primary and 3,824 metastatic cancer genomes of the PCAWG consortium and the HMF cohort shows that practically all signatures operate dynamically in response to various genomic and epigenomic states. The analysis pins differential spectra of UV mutagenesis found in active and inactive chromatin to global genome nucleotide excision repair. TensorSignatures accurately characterises transcription-associated mutagenesis, which is detected in 7 different cancer types. The algorithm also extracts distinct signatures of replication- and double strand break repair-driven mutagenesis by APOBEC3A and 3B with differential numbers and length of mutation clusters. As a fourth example, TensorSignatures reproduces a signature of somatic hypermutation generating highly clustered variants around the transcription start sites of active genes in lymphoid leukaemia, distinct from a more general and less clustered signature of Polη-driven translesion synthesis found in a broad range of cancer types. Finally, I demonstrate TensorSignatures' utility by applying it to multiple datasets in various collaboration projects.

Taken together, TensorSignatures adds great detail and refines mutational signature analysis by jointly learning mutation patterns and their genomic determinants. This sheds light on the manifold influences that underlie mutagenesis and helps to pinpoint mutagenic influences which cannot easily be distinguished based on the mutation spectra alone. As mutational signature analysis is an essential element of the cancer genome analysis toolkit, TensorSignatures may help make the growing catalogues of mutational signatures more insightful by highlighting mutagenic mechanisms, or hypotheses thereof, to be investigated in greater depth.

PETE: The Visilia Oaks and our 240-pound catcher, Jeremy Brown, who, as you know, is scared to run to second base. This was in a game six week ago. This guy's gonna' start with a fastball. Jeremy's gonna take him to deep center. Here's what's interesting, because Jeremy's gonna do what he never does. He's gonna go for it. He's gonna round first and he's gonna go for it. Okay? This is all of Jeremy's nightmares coming to life.

BILLY: Aw, they're laughing at him.

PETE: And Jeremy's about to find out why... Jeremy's about to realise that the ball went 60 feet over the fence. He hit a home run and didn't even realise it.

BILLY: How can you not be romantic about baseball ?

PETE: It's a metaphor.

BILLY: I know it's a metaphor.

from Moneyball

Acknowledgements

Throughout my PhD studies I have received a great deal of support and assistance. I would first like to thank my supervisor, Dr. Moritz Gerstung, whose expertise was invaluable in formulating the research topic and methodology in particular. It was your vision, leadership and relentlessness that kept me motivated throughout the last four years. Thank you for being a patient, but also demanding teacher, I could not have chosen a better one!

I would like to acknowledge my colleagues from the Gerstung Group, including Nadezda Volkova, Santiago Gonzalez, Sarah Killcoyne, Yu Fu, Rui Costa, Alexander Jung, José Almeida, Artem Lomakin and Stefan Dentro. Nadezda, I want to especially thank your for you excellent organisation skills and your generous help in proofreading my thesis.

I like to appreciate my collaborators from the IRB Barcelona and the UMC Utrecht, particularly Nuria Lopéz, Oriol Pich, Arne van Hoeck and Edwin Cuppen, for providing me the opportunity to learn from your expertise and allowing me to work with your data.

I would also like to thank my fellow Predocs, particularly Sushmita Sridhar and Hannah Currant, who accompanied me on this journey, and became two of my dearest friends along the way.

In addition, I would like to thank my parents and my dear girlfriend Chloé Bleuez. You are always there for me, care about me, overlook my mistakes and endure me even if I do you wrong.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Cancer is the second leading cause of death globally, responsible for approximately 9.7 million deaths in 2018[1]. While we nowadays take the molecular foundations of carcinogenesis for granted, it took almost 200 years to arrive at these insights, and began with biological concepts established by the nineteenth-century giants Mendel and Darwin, who outlined the foundation for genetics and evolution.

Gregor Mendel established the basic rules of genetics by tracking the breeding of pea plants in 1860. He found out that genetic information can be understood as a collection information packets, in today's language genes, that define distinct physical traits in an organism. Since genetic information is passed in its entirety from parent to offspring, many organisms store their genetic information in twofold redundancy, allowing them to preserve the genetic information from both parents. Different versions of a gene are called alleles. For example, if an organism carries two identical versions of a gene, it is said to be homozygous, while an individual with two distinct copies is heterozygous with respect to the gene (Weinberg, 2013).

In the early twentieth century, it seemed that genetic variability must have been established at the evolutionary beginnings of a species, but this view was challenged when researchers found out that genetic information is corruptible. Mutations can change the information content of a gene, thus interchanging one allele into another, or creating completely novel gene versions. In some sense, mutations provide species the means to tinker with their genomes, giving organisms the opportunity to continually improve their phenotype, i.e. the composite of observable characteristics of an organism, due to new versions of genes. This results in a diversification of alleles over the evolutionary history of a species, such that older species harbour more distinct alleles in comparison to more recent ones (Weinberg, 2013).

[1] According to the World Health Organisation (2018).

However, not all mutations turn out to be advantageous, in fact, most of them have adverse effects, some mutations may be even deleterious. Charles Darwin recognised the interplay between evolutionary forces driving the formation of mutations and environmental constraints: Nature *naturally select* alleles conferring favourable phenotypes, while alleles with unfavourable effects are continually discarded (Weinberg, 2013). This insight has profound consequences to conceptualise cancer, because tumours take advantage of mutations to ensure their persistence, thereby exploiting evolutionary mechanisms against their host organisms.

Watson and Crick inaugurate the biological revolution of the twentieth century

Although the concepts of Mendel and Darwin are probably the most pervasive in biology, it took almost 100 years to resolve the structure of the molecule that embodies the genetic constitution of a cell. The time had come in 1953, when Watson and Crick announced the discovery of the deoxyribonucleic acid (DNA) double helix, and humbly[2] acknowledged that "*It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.*" (Watson and Crick, 1953), thereby providing the molecular and mechanistic bases that explain evolution, enable modern genetics, and would launch the biological revolution of the twentieth century.

Initially, a large proportion of DNA had no known biological function, and was therefore termed as "junk DNA" (Gregory, 2011). In fact, only 1.5 % of a mammalian genome is dedicated to encode protein sequences, and another 2 % is important to regulate gene expression and other functions (Lander, 2011; Weinberg, 2013). Consequently, since mutations affect random locations in the genome, most genomic alterations have no or very little effect on the cellular or organismic phenotype. For this reason, mutations in non-encoding genomic regions are said to be silent or neutral, because from an evolutionary point of view, they confer neither advantages nor disadvantages. The lack of phenotypical implications made it impossible for early geneticists to detect neutral mutations, and it only became feasible after the first DNA sequencing techniques were invented. These revealed that each human genome carries its own unique collection of genomically silent mutations, often comprising millions of so-called genetic polymorphisms (Lander, 2011).

While the configuration of genetic polymorphisms is transmitted from parent to offspring, it is important to distinguish them from germline or somatic mutations. Germline mutations affect every cell of the offspring by striking the genome of a sperm or egg, or their precursor

[2]It should be mentioned that Watson and Crick made use of Rosalind Franklin's crystallographic evidence, which was shown to them without her consent by her colleague Maurice Wilkins, to confirm their guess about the double helical DNA structure.

stem cells within the gonads, respectively. On the other hand, somatic mutations arise everywhere else except from the germline, and only affect particular cells. Some somatic mutations may *drive* cancer formation, as they change the properties of the affected cell and all of its descendants (clones) within a tissue (Weinberg, 2013). Such mutations affect genomic regions that encode genes with oncogenic or tumor suppressive potential. Over time, the descendants of this clone will acquire further mutations, eventually equipping some of the daughter cells with additional invasive properties, allowing them to turn into to malignant tumours.

In particular, the cells involved in tumour formation develop the following properties: (1) cell growth and division in absence of proper signals, (2) continuous growth and division even in presence of contrary signals, (3) avoidance of apoptosis, (4) limitless number of cell divisions, (5) promotion of blood vessel construction, and (6) invasion of tissue and formation of metastases (Hanahan and Weinberg, 2000, 2011). To acquire these capabilities, normal cells undergo a multistage transformation in which a pre-cancerous lesion progresses to a malignant tumour. This process results from an interaction between an individual's genetic predisposition and various environmental factors, which induce genomic alterations that change the normal cellular program.

The origins of cancer

While the principles established by Mendel and Darwin are essential to conceptualise the disease, they tell us little about how mutations actually emerge and how cellular factors influence mutagenesis. The first evidence indicating that mutations are not purely spontaneous and random was given by the fact that the incidence of many cancers vary by country. Although heredity or environment could account for this variation, epidemiological studies found lifestyle factors to be the main determinants of the country-by-country variation in cancer incidence. This idea was supported by laboratory research that narrowed down chemical and physical agents such as tobacco, coal dust and X-rays as potent carcinogens. In addition, viruses were found to cause leukemias in chickens, raising the possibility that cancer is an infectious disease (Weinberg, 2013). However, in 1975 Ames provided strong experimental support for the hypothesis that carcinogenic substances induce cancer by mutating genes, indicating that a large fraction of cancers may be attributed to the exposure to a wide range of substances (Ames et al., 1975). Over time, researchers characterised the impact of different mutagens on DNA, which revealed that many carcinogens introduce specific mutations.

With the advent of next-generation sequencing techniques, it became feasible to detect the entirety of somatic mutations within cancer genomes, including both driver and neutral

mutations. While the analysis of driver mutations leads to the identification of genes with oncogenic or tumor suppressive potential, it is less obvious what value is gained by analysing neutral mutations that lack phenotypic consequences. However, although neutral mutations do not promote tumour progression, they still resemble a genomic imprint of carcinogenic exposure due to an endogenous or exogenous mutagenic source, thereby conveying great information about the etiologic (causative) mechanisms of cancer development (Helleday et al., 2014).

While the analysis of sequenced cancer genomes largely confirmed that cancer types such as melanomas or lung cancers exhibit predominantly mutations that one would expect considering the mutagens they have been exposed to, it surprisingly also revealed that mutagenesis does not occur uniformly on the genome, indicating that the genomic organisation and biological processes involving DNA affect mutagenesis (Schuster-Böckler and Lehner, 2012). At the same time, the first attempts to systematically characterise the mutational patterns of cancers were made by applying mathematical approaches to mutation count data (Alexandrov et al., 2013a). Over time, these efforts provided more than 50 different single base substitution (SBS) patterns, so-called mutational signatures, indicative for a range of endogenous mutational processes, as well as genetically acquired hypermutation and exogenous mutagen exposures (Alexandrov et al., 2018).

This overview of carcinogenesis illustrates the complexity of the topic and shall prepare the reader for the intricacy that has yet to be explored when mutagenesis is studied on a whole genome level. So far, a number of studies have analysed cancer genomes to extract mutational processes using computational pattern recognition algorithms such as non-negative matrix factorisation (NMF) over catalogues of single nucleotide variants (SNVs) and other mutation types. However, these mutational signatures characterise mutational processes only in terms of their mutational imprint but fail to take into account the complexity of the genome, or reflect the multitude of different mutation types caused by a single process. The goal of this research was to address these shortcomings by developing a novel method capable of incorporating genomic determinants such as the local genome structure, or fundamental DNA involving processes including transcription and replication. To motivate this, I will start by reviewing the fundamental properties of DNA and its nuclear organisation (Sec. 1.1), as well as transcription and replication (Sec. 1.2). In Sec. 1.3, I will provide an overview of endogenous and exogenous mutational processes, and discuss in Sec. 1.4 the mechanisms cells have evolved to repair them. Sec. 1.5 discusses mutagenesis in the context of the genome, and Sec. 1.6 introduces the computational method NMF, which is at the heart of mutational signature analysis, and summarises the insights that have been drawn from this

methodology. Finally, in the last section of this chapter (Sec. 1.7), I will briefly layout the structure of this thesis.

1.1 Eukaryotic DNA structure and organisation

The fact that DNA plays a central role in various important cellular processes and consists of structurally and functionally distinct compartments introduces complexity to mutagenesis and DNA repair. To appreciate this, I will review in this section the basic (Sec. 1.1.1) and higher order structure (Sec. 1.1.2) of the molecule.

1.1.1 Nucleotides are the building blocks of the DNA double helix

Nucleotides make up the basic building blocks of DNA and are crucial to mediate the structure and functionality of the molecule. They comprise three components: a five carbon sugar (desoxyribose), a phosphate group, and one of four possible nitrogenous bases: the purines adenine (A) and guanine (G), and the pyrimidines cytosine (C) and thymine (T). The sugar component misses a hydroxyl group at its 2'-carbon, which is normally present in ribose, attaches to the base component via an N-glycosidic bond at its 1' carbon, and forms a covalent bond with a phosphate group to connect to neighbouring nucleotides via its 3' or 5' carbon (Fig. 1.1a, Alberts et al. (2007)). A single DNA strand is formed by connecting a sequence of nucleotides. Here, the carbon numbering is key to describing the intrinsic directionality of the polymer which is in standard orientation 5' to 3' (Fig. 1.1b).

Structural properties of the DNA double helix enable protein interactions

Two DNA strands, when aligned in opposite directionality, may interact through non-covalent hydrogen bonds between their bases. Hydrogen bonding contributes to the specificity of base pairing, because T preferentially pairs with A, and C preferentially pairs with G through two and three hydrogen bonds, respectively (Fig. 1.1b). The geometries of the A:T or T:A and G:C or C:G base pairs are the same, allowing for symmetry and base stacking, and enabling two complementary strands to intertwine, thereby forming the characteristic double helix. The helical structure of double-stranded DNA (dsDNA) features structural properties important for DNA interactions with proteins. Particularly, it constitutes a major groove which allows proteins with fitting domains to bind while distinguishing between A:T and T:A base pairs and between G:C and C:G base pairs, thus enabling sequence specific protein interactions. On the other hand, the minor groove presents less information, allowing bound

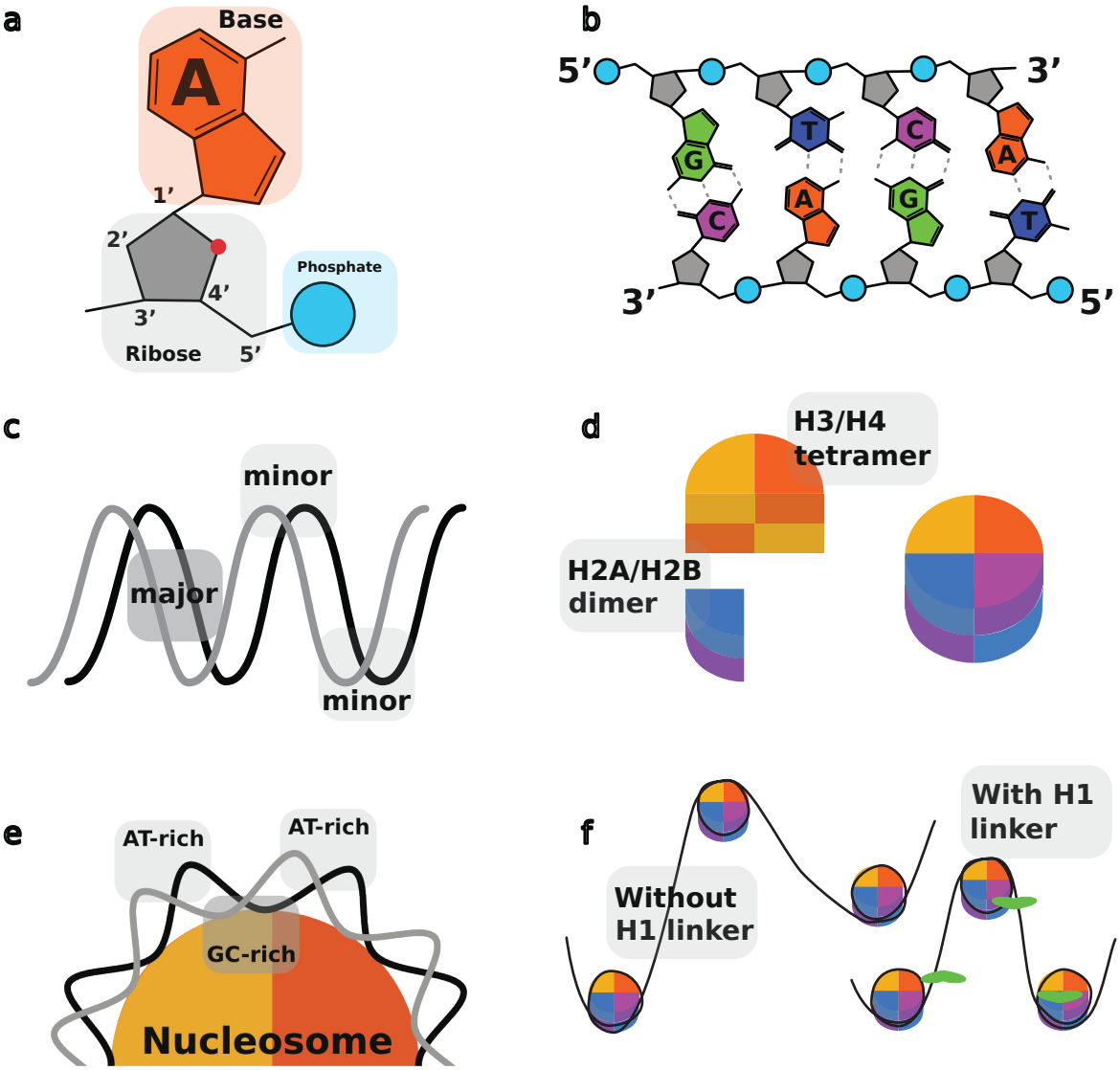


Fig. 1.1 **DNA structure and organisation. a**. Nucleotides comprise three distinct structural components: a ribose, a base and a phosphate group. **b**. Two antiparallel DNA strand align to each other by forming hydrogen bonds (dashed lines) between complementary nucleotides. **c**. The DNA double helix features DNA-protein interaction by exposing a major and a minor groove. **d**. A nucleosome is composed of a H3/H4 tetramer and two H2A/H2B dimers. **e**. Nucleosomal DNA binding is facilitated by minor-groove interactions between the histone protein complex and DNA. **f**. H1 linker histones enable further compaction and promote higher order DNA structure (not shown).

proteins to only differentiate between A:T/T:A and G:C/C:G base pairs (Fig. 1.1c, Alberts et al. (2007)).

1.1.2 Nucleosomes establish several layers of DNA compaction

The total length of the human genome corresponds to a 3×10^9 long sequence of base pairs. This corresponds to a 2 m long unrolled string of DNA, whereas an average human nucleus has a diameter of only 10 µm, suggesting that cells must have evolved efficient ways to compress the linear molecule.

Histone proteins form the core of nucleosomes

The first stage of compaction is achieved by wrapping 147 bp DNA 1.7× around disk-shaped octameric histone protein complexes (each DNA-protein complex contains two copies of H2A, H2B, H3 and H4, Fig. 1.1d, Luger et al. (1997)). Histone proteins are highly conserved across eukaryotes, positively charged due to a high of amount basic amino acids such as arginine and lysine, and exhibit unstructured N-terminal tails that stick out of the DNA-protein complex to enable regulation. The combination of histone proteins and associated DNA is termed a nucleosome and represents the most basic structural unit of DNA packaging (Alberts et al., 2007).

Histone proteins bind the phosphate back-bone and the minor groove of DNA. Although these interactions are largely non-sequence-specific, there are certain sequence compositions that facilitate DNA-histone interactions by inducing favourable configurations of the molecule (Segal et al., 2006). Particularly well-suited sequences are alternating GC and AT-rich minor groove sequences (Fig. 1.1e), because GC-rich minor grooves bend DNA towards the major groove, and thus away from the histone, while AT-rich minor grooves bend the molecule towards the minor groove and the nucleosome (Kaplan et al., 2009; Segal et al., 2006). In this way, a single histone octamer forms approximately 40 hydrogen bonds with DNA.

Linker histone achieve further compaction and promote higher order nucleosomal organisation

Although nucleosomal DNA packaging compacts DNA by roughly ∼5.5-fold, further compression is required to fit the genome into the nucleus, which is achieved by assembling H1 linker histones to the DNA-protein complexes. In contrast to other histone proteins, H1 does not locate to the nucleosomal core, but rather sits on top of the structure by binding to the nucleosome bound DNA midpoint (dyad), thereby compacting the molecule up to 40-fold (Fig. 1.1f, Ramakrishnan et al. (1993)). Linker histone binding also keeps in place the wrapped DNA, facilitating the formation of higher-order structures involving many nucleosomes into solenoid or zig-zag 30 nm fibres (Robinson et al., 2006; Schalch et al., 2005). It is believed

that 30 nm fibres loop around a proteinaceous structure called nuclear matrix to form whole chromosomes (Alberts et al., 2007).

1.2 Genome regulation and activity

DNA condensation is crucial to store DNA compactly, but also makes large parts of the genome inaccessible to DNA-binding proteins, which represses many obligatory cellular processes such as transcription or replication. For this reason, cells control structural and functional properties of DNA by epigenetic modifications, which do not alter the sequence information, but rather the state of DNA associated proteins or nucleotides to dynamically modulate the accessibility of specific genomic regions.

1.2.1 Epigenetic regulation of DNA accessibility

The most basic level of DNA regulation is mediated by the way DNA is wrapped around the histone core. DNA accessibility is highest at the DNA entry points to the nucleosome, and declines gradually towards the dyad. To understand this, consider that DNA segments located at nucleosomal entry points are much less tightly bound, enabling the DNA to temporarily dissociate from the protein complex and expose their sequence to DNA-binding proteins. In contrast, DNA proximal to the dyad is surrounded by tightly bound DNA on both sides, making the latter event less likely to occur (Klemm et al., 2019; Li et al., 2005).

Histone remodelling complexes modulate DNA accesibility by sliding DNA around nucleosomes and exchanging histone subunits

One important class of proteins dedicated to alter the accessibility of DNA are nucleosome remodelling complexes, which utilise adenosine triphosphate (ATP) to slide DNA around the disk-shaped protein complex, or catalyse the removal or the exchange of histone subunits, often with non-standard histone isoforms that mark functionally distinct genomic regions (Saha et al., 2006). For example, histone variant CENP-A is found at centromers and may interact with microtubules during anaphase of mitosis (Chueh et al., 2005), H3.3 marks sites of transcription (Chen et al., 2013), H2A.X indicates double strand breaks in DNA (Ayoub et al., 2008), and H2A.Z appears to locate at boundary regions to nucleosome free regions (Hatch and Bonner, 1990).

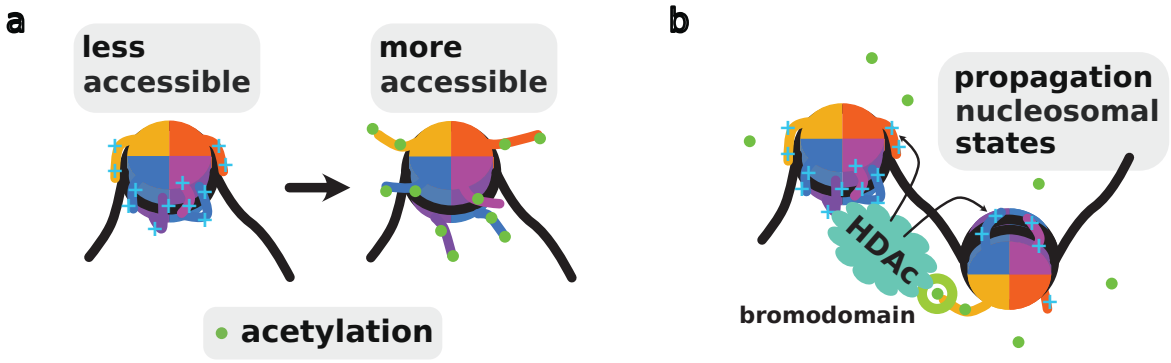


Fig. 1.2 **Epigenetic modulation of DNA accessibility. a**. Acetylation of positively charged amino acids in N-terminal histone tails reduces the nucleosomal affinity for bound DNA. **b**. An example of epigenetic regulation of DNA accessibility via postranslational modification of histone tails (see text).

Postranslational modification of histone tails modulate DNA access by structural perturbations and secondary effector proteins

Another form of epigenetic modulation is mediated by histone tails, which are unstructured and positively charged, extend outside of the nucleosome, and are thus accessible to enzymes that add or remove diverse molecules ranging from small acetyl groups to larger proteins such as ubiquitin. The two most important histone modifications are acetylations and deacetylations of lysins by histone acetyltransferases (HATs) and histone deacetylases (HDACs), and methylations and demethylations of lysins or arginines by histone methyltransferases (HMTs) and histone demethylases (HDMes) (Lee and Workman, 2007; Milazzo et al., 2020).

It is thought that histone acetylation makes DNA more accessible by neutralising the (positive) amino group of lysins, thus reducing the affinity of histone N-terminal ends for the negatively charged DNA (Fig. 1.2a). Acetylation also interferes with 30 nm fibre formation, because the establishment of this higher order structure partly depends on interactions mediated by positively charged histone tails and adjacent nucleosomes (Eberharter and Becker, 2002). While acetylations induce direct structural perturbations, other histone modifications like methylations, but also acetylations, mediate their effects via effector proteins equipped with domains that recognise certain histone modifications. Examples for such domains are bromodomains which recognise acetylated nucleosomes, or TUDOR (Huang et al., 2006; Lee et al., 2008), PHD finger (Iwase et al., 2007), and chromodomains that direct effector proteins to methylated histone tails (Min et al., 2003).

The association of flexible histone extensions to effector proteins with histone modification recognising domains allow the propagation of histone states, enabling cells to quickly alter functional and structural properties of *specific* genomic segments (Fig. 1.2b). For

instance, a HDAc with a bromodomain may promote genome silencing in active genomic regions. The bromodomain enables the enzyme to target acetylated histone tails, which are mainly found in transcribed genomic regions. Once bound to such a nucleosome, the HDAcs may deacetylate the N-terminal tails of the bound and adjacent nucleosomes to promote chromatin condensation (Fig. 1.2b).

Histone codes mark functionally distinct genomic regions

The discovery of histone tail modifications led to the proposition of the histone code hypothesis, stating that specific posttranslational modifications serve to recruit distinct proteins with matching recognition domains, which then alter the chromatin structure to promote certain downstream events (Jenuwein, 2001; Strahl and Allis, 2000). Today, this theory is considered certain, particularly the methylation of H3K4 and H3K36 correlates with transcriptional activation, while demethylation of the former silences genomic regions, transcriptional repression is associated with H3K9 and H3K27 methylation (Hublitz et al., 2009), and trimethylation of H3K9 indicates constitutive heterochromatin (Hyun et al., 2017).

Cytosine methylations regulate the transcription rate of genes and affect local chromatin structure via methyl-CpG-binding domain proteins

Cytosine methylation at CpG sites alter the accessibility to regulatory proteins, thus influencing the transcription rate of genes. Interestingly, this may either result in the activation or repression of genes depending on the location of the methylation within the gene. Methylated cytosines at gene regulatory regions may also remodel the local chromatin structure by recruiting methyl-CpG-binding domain (MBD) proteins, which interact with nucleosome remodelling complexes and HDAcs, leading to gene silencing (Bhattacharya et al., 1999).

1.2.2 Mechanics of transcription

Transcription is the process by which the information in DNA is copied into a new molecule of ribonucleic acid (RNA). While there are three different RNA polymerases in eukaryotes to generate various RNA products, I will focus in this section on RNA polymerase II (RNA Pol II). This polymerase exclusively transcribes messenger RNA (mRNA), which is subsequently translated into proteins. In eukaryotes, transcription is naturally repressed due to the nucleosomal organisation of DNA, making transcriptional activation the dominant mode of regulation.

Proximal and distal DNA regulatory elements control transcription

Transcriptionally active regions constitute a promoter, i.e. a site that directs RNA Pol II to bind a gene. The core promoter is a $\sim$60 bp long sequence encompassing a transcription start site (TSS), and up to four general transcription factor (TF) binding sites such as the TATA box, indicative for how efficiently RNA Pol II is recruited (Alberts et al., 2007). The proximal promotor subsumes the core promotor plus regulatory elements 200-400 bp upstream of it. The size of a full promotor ranges from 400 to 10 kbp, contains the proximal promotor and long range promotors, including enhancers or insulators (Fig. 1.3a).

Enhancers may be located upstream, downstream, or within transcribed regions (generally in introns). While some details of enhancer actions are still elusive, it is established that enhancer RNA (eRNA) is required for looping, which brings enhancer and proximal promoter elements together (Kim et al., 2010; Wang et al., 2011a). Enhancer recruit many of the same factors as the promoter (e.g. nucleosome remodelling complexes) and provide in this way activation reinforcement. In contrast, insulators establish chromatin boundaries by blocking the action of enhancers through competing looping mechanisms, or by forming chromatin boundaries by stopping the propagation of active histone states (Gaszner and Felsenfeld, 2006; West et al., 2002).

Transcription comprises three distinct stages: initiation, elongation and termination

Transcription initiation starts with the assembly of the pre-initiation complex constituting several (general) TFs, which bind to recognition sites such as the TATA-box, and RNA Pol II at the proximal promotor (Fig. 1.3b). Subsequently, TFIIH promotes DNA opening by translocating DNA towards the TSS, and initiates promotor escape by phosphorylating serine-5 of RNA Pol II C-terminal domain (CTD) (Goodrich and Tjian, 1994), which is the unstructured "tail" of the enzyme, consisting of 52 heptapeptide repeats (Hsin and Manley, 2012). This starts transcription by freeing RNA Pol II from the Mediator (Fig. 1.3b), a large multi protein complex that is thought to bind unphosphorylated CTD to elaborate RNA Pol II's surface area, enabling the transcription machinery to interact with a larger number of TFs, which often locate to distal enhancer elements. Serine-5 phosphorylation also recruits capping enzymes targeting the 5' end of nascent mRNA (Spangler et al., 2001). The 5' cap protects mRNA from exonucleases and helps the translation machinery to recognise transcripts (Fig. 1.3b, Alberts et al. (2007)).

Many promoters feature a promoter proximal pausing event mediated by negative elongation factor (NELF) and DRB sensitive factor (DSIF) which arrest RNA Pol II shortly after promotor escape ($\sim$25-60 bp, Adelman and Lis (2012)). The positive transcription

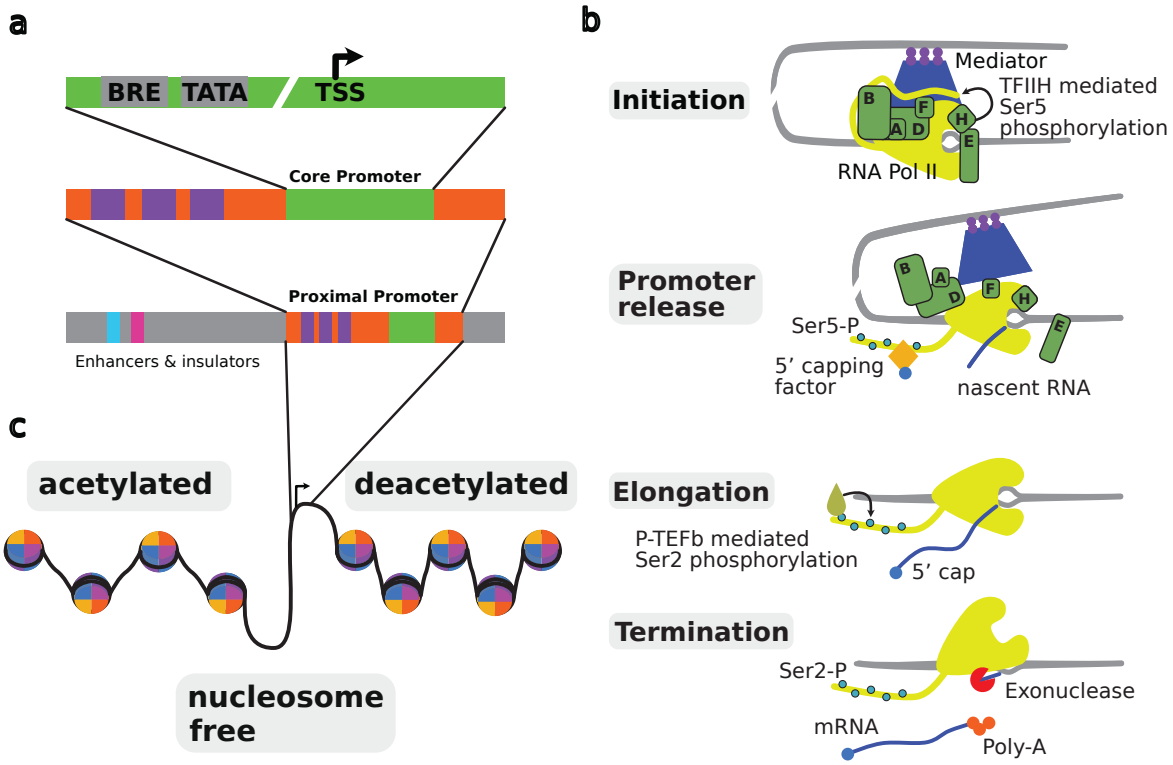


Fig. 1.3 **Eukaryotic transcription. a**. Structure and components of an eukaryotic promotor. **b**. The three phases of transcription constitute initiation, elongation and termination (see main text for details). **c**. DNA organisation at a transcribed loci.

elongation factor (P-TEFb) initiates processive transcription by phosphorylating NELF and DSIF, causing the former to leave the transcription machinery. During transcription elongation, P-TEFb also phosphorylates serine-2 of RNA Pol II CTD (Fig. 1.3b), which recruits RNA splicing enzymes and the 3' processing machinery necessary to add the Poly A tail to emerging mRNA transcripts. At the same time, the amount of serine-5 CTD phosphorylation decreases, causing mRNA 5' end modifying enzymes to leave the CTD (Narita et al., 2003).

Transcription termination does not rely on RNA Pol II, but is initiated by the polyA-signal of the mRNA transcript, which is usually located shortly after the stop codon. This recognition sequence is detected and bound by the 3' processing machinery located at the RNA Pol II CTD, which cleaves the nascent mRNA and adds an untemplated poly-A tail to the transcript, thus protecting the newly generated mRNA molecule from 3' degradation (Guhaniyogi and Brewer, 2001). At this stage, RNA Pol II continues to elongate the cleaved RNA molecule, which is due the lack of a 5' cap prone to degradation by exonucleases. This enzyme "chews off" the RNA from the active site of RNA Pol II, resulting in reannealing of the coding and template strand DNA (Fig. 1.3b), thus removing the substrate for RNA Pol II, which is unable to bind DNA again without the help of general TFs (Kim et al., 2004).

The chromatin structure of actively transcribed genomic regions

There are strong links between transcription regulation and the epigenetic configuration of transcribed loci: Promoter and TSS are usually nucleosome free, enabling TFs and RNA Pol II to bind and initiate transcription, while regions upstream of promotors are acetylated, characteristic of accessible chromatin. However, a somewhat surprising observation is that coding regions, downstream from gene promotors, are usually in deacytelated form, raising the question how RNA Pol II transcribes through poorly accessible chromatin (Fig. 1.3c, Owen-Hughes and Gkikopoulos (2012); van Steensel and Furlong (2019)).

As it turns out, phosphorylated RNA Pol II CTD itself regulates transcribed loci by recruiting HMT SET2, which methylates H3K26. This recruits HDACs to remove acetyl groups from histone tails, resulting in tightly organised chromatin (Venkatesh and Workman, 2013). To nevertheless transcribe through the condensed chromatin, the transcription machinery makes use of the facilitates chromatin transcription complex (FACT), which removes a H2A/H2B dimer from a nucleosome, leaving behind a histone hexamer (hexasome) that is easier to transcribe through. FACT in conjunction with transcription elongation factor SPT6 restore the nucleosomal structure by subsequently reconstituting removed histone parts (Venkatesh and Workman, 2013).

1.2.3 DNA replication

Cells employ DNA replication in mitosis or meiosis to produce two identical replicas of the DNA molecule.

DNA synthesis at a replication fork requires three different polymerases

The replication fork is a structure that forms within dsDNA during DNA replication. Here, the antiparallel strands are split into single-stranded DNA (ssDNA), ready to be processed by DNA polymerases. In fact, eukaryotic replication involves three different DNA polymerases. DNA polymerase ε (Pol ε) and DNA polymerase δ (Pol δ) copy leading and lagging strand, respectively (Alberts et al., 2007). While the leading strand DNA polymerase always copies DNA in the same direction as the movement of the replication fork, lagging strand DNA synthesis occurs in the opposite direction, making it necessary for Pol δ to replicate to a stopping point, and then start again at a position closer to the replication fork. In contrast, DNA polymerase α (Primase) synthesises 6-8 bp RNA primers at replication forks to provide primer:template junctions (PTJs) which serve as substrates for replicative polymerases (Frick and Richardson, 2001).

Temporal properties of the DNA synthesis reaction and the structure of DNA polymerases ensure high-fidelity DNA synthesis

DNA synthesis requires a PTJ, i.e. a primer with a free 3'OH annealed to a longer stretch of template ssDNA, and all four types of deoxyribonucleotide triphosphates (dNTPs). DNA polymerases recognise PTJs as substrates and extend the primer 3'OH by adding dNTPs matching the template sequence. It is important to realise that DNA polymerases cannot distinguish a correct from an incorrect dNTP by interacting with the incoming dNTP alone. When a matching base enters the active site, it base pairs the template nucleotide, which positions the α-phosphate group of the incoming dNTP close to the primer 3' OH, facilitating quick catalysis (Alberts et al., 2007). However, in the absence of a correct base pair, the α-phosphate of the incoming dNTP is not in close proximity to the primer 3' OH, resulting in much slower catalysis and enabling the incorrect dNTP to fall out of the active site again (Alberts et al., 2007).

While temporal regulation plays an important role in accurately distinguishing between similar dNTPs, structural features of DNA polymerases enforce additional spatial constraints to ensure high fidelity DNA synthesis. To understand this, recall that replicative DNA polymerases have the shape of a hand (Doublié and Zahn, 2014), with the palm binding the PTJ through interactions with the phosphate backbone, and non-sequence, but Watson-Crick specific minor groove interactions, enabling the enzyme to detect whether the newly synthesised dsDNA displays the hydrogen donor/acceptor pattern of matching base pairs. In contrast, the finger domains of DNA polymerases properly orient participating reactants, particularly important is the so called O-helix, which clamps the incoming residue so that water cannot enter the active site, and only the nucleophilic attack of the primer 3'-OH can occur (Hübscher et al., 2002; Johnson, 1993; Joyce and Steitz, 1994).

The exonuclease domain of replicative DNA polymerases increase fidelity of DNA synthesis

Temporal properties of the chemical reaction and structural features of DNA polymerases enable cells to replicate DNA with error rates of 1 mistake per 10^5 bp synthesised, an unacceptably high rate considering that the human genome has the size of 3×10^9 bp[3]. The mistakes escaping these precautionary measures are mostly transitions (A>G or C>G), and due to the chemistry of nucleotides, which may spontaneously switch from keto to enol form (tautomery), albeit these transitions are rare and last only very brief periods of time (Wang et al., 2011b). To understand this, consider that a guanine in its common *keto* state

[3]This corresponds to 3,000 errors per round of replication (Kunkel, 2009)!

pairs with cytosine, but rather interacts with thymidine in its *enol* form due to an altered pattern of exposed hydrogen acceptors/donors. In this way, temporarily present enols in the template DNA may cause the incorporation of wrong dNTPs. However, as soon as an enol switches back to its energetically more favourable keto form, the wrongly incorporated base destabilises the PTJ enough to create a small region of ssDNA at the 3' end, for which the DNA polymerase proofreading exonuclease domain has a 10-fold higher affinity in comparison to the DNA polymerase active site. The proofreading exonuclease domain will remove a few nucleotides before the PTJ, allowing to restore the dsDNA and enabling DNA polymerase to proceed with DNA synthesis (Swan et al., 2009). Overall, the presence of the 3' to 5' exonuclease domain increases the accuracy of replicative polymerases by a 100-fold so that one round replication would leave an imprint of 30 mutations, which is still to high, but is addressed by additional DNA repair mechanism discussed in later sections (Kunkel, 2009).

Other enzymes at the replication fork involve helicases, replication protein A, and topoisomerases

Although DNA polymerases perform most of the work during DNA replication, cells also rely on various other enzymes to ensure flawless proceeding of the DNA copying process. For example, one problem eukaryotic cells face is the burden of copying vast amounts of DNA due to their large genome size, making processive (fast) DNA replication indispensable. To accomplish this, cells employ replicative helicases, which are hexameric ring-shaped molecules that unwind DNA at the replication fork under the consumption of ATP. The molecular structure of DNA helicase suggests that the enzyme encircles the leading strand in eukaryotes (Enemark and Joshua-Tor, 2006), and promotes movement of the replication fork by a "paddling" mechanism that pulls the bound ssDNA through the enzyme's central hole, thereby displacing both DNA strands and enabling up to 6× faster replication (Ha, 2007).

Pol δ synthesis of the lagging strand is discontinuous, and occurs in the opposite direction to replication fork progression, causing the latter strand to remain for a longer period in fragile ssDNA configuration (Alberts et al., 2007). To stabilise the DNA during this critical period, and to prevent premature reannealing of both DNA strands, cells employ replication protein A (RPA), which readily binds ssDNA, but not dsDNA, in a non sequence specific manner. Notably, RPA binding is cooperative, meaning that if one binds it is easier for a second to bind as well, and it does not prohibit DNA polymerases to replicate over its bound DNA (Bae et al., 2001).

Apart from RPA, additional enzymes are required to finalise the synthesis of the lagging strand. To remove the RNA primer at the beginning of each Okazaki fragment, FEN1 de-

grades most of the primer except the last rNMP, for which an additional 5' to 3' exonuclease is recruited (Cerritelli and Crouch, 2009). The resulting gap is then filled by a DNA polymerase, and a specialised DNA ligase which links the newly generated DNA to following the Okazaki fragment (Alberts et al., 2007).

Unwinding the double helix during DNA replication results in positive supercoiling, which refers to the overwound DNA that has yet to be processed by the replication machinery. This could lead to a stalling replication fork, because a surplus of DNA twists increases the strain energy stored in the molecule, making it more difficult for DNA helicase to proceed with unwinding the DNA. To overcome this issue, cells employ type I DNA topoisomerases[4], which cut a single strand of dsDNA and relax supercoils by pulling the other strand through the incision such that the molecule unwinds (Champoux, 2001; Wang, 2002).

Random redistribution of parental DNA histone proteins to both daughter DNA ensures proper propagation of epigenetic states

Although replication is primarily associated with the duplication of DNA, it is also the time when the chromosome structure of the genome is established and associated proteins are copied. Importantly, the chromatin structure ahead of the replication fork needs to be disassembled as the replication machinery moves along the genome, as well as put back together in the *right configuration* on both daughter DNA molecules. Nucleosome assembly typically proceeds with the following steps: first, H2A:H2B and H3:H4 dimer assemble, followed by H3:H4 tetramer formation. Histone chaperones escort H3:H4 tetramers to DNA, where the protein complex binds DNA by wrapping the molecule around it such that the dyad and both ends of DNA (the entry and exit point) are bound to the tetramer. Two H2A:H2B dimers subsequently join the $2 \times$ H3:H4 DNA complex, thereby completing the histone octamer (Dennehey and Tyler, 2014). As the replication fork passes, nucleosomes of the parental DNA are disassembled, and the pool of freed H3:H4 tetramers are randomly distributed to both daughter DNA molecules, enabling to rebuild nucleosomes from this point. Although the newly generated DNA also comprises some completely new nucleosomes, marks present on redistributed histones suffice to reconstitute the original epigenetic state, due to histone modifying enzymes capable of propagating the signals present on parental histones (Sec. 1.2.1) (Ransom et al., 2010; Serra-Cardona and Zhang, 2018).

[4]There are also type II topoisomerases, but these are less relevant for DNA replication.

Replication starts at different times at various loci of the genome

The discussion of DNA replication so far naively suggests to initiate a replication fork at the beginning of a chromosome, and to replicate its DNA until DNA polymerase reaches to its end. This is however not the case, as eukaryotic DNA replication starts at multiple sites of the genome, termed origin of replication (ORI), where DNA is initially unwound and DNA synthesis is started. This brings the advantage of parallelism, which enables cells to replicate the genome much faster.

The human genome contains approximately 30,000 ORI, initiating at characteristic times, with some firing early and others rather late in S-phase. Although there is little known about the mechanisms orchestrating the temporal pattern or its significance, it is thought that the timing pattern of ORI initiation may help to ensure that replication finishes even in presence of DNA damage. Replication forks may stall upon encountering DNA alterations. When DNA damage(s) block the progression of two convergent replication forks stemming from two early-firing origins, then a late-firing origin from the unreplicated region in between may be activated to close the gap between stalled adjacent forks (Yekezare et al., 2013).

1.3 DNA damage

The information encoded in the DNA sequence of each organism ensures its survival and must therefore be passed on correctly to the next generation. However, not only the mere fact that the DNA molecule has to survive within a physiological environment, but also the exposure to exogenous factors may harm the molecule in various ways. In this section, I will outline in what ways endogenous (Sec. 1.3.1) and exogenous (Sec. 1.3.2) sources may harm genetic information.

1.3.1 Endogeneous sources of DNA damage

Endogenous DNA damage arises mainly due to hydrolytic and oxidative reactions of nucleobases with water and reactive oxygen substances (ROS), which are both naturally present within cells.

Spontaneous and enzymatic hydrolytic deamination of cytosines

Although DNA is a chemically stable molecule, the presence of water in cells causes constant hydrolysis of nucleobases. For example, hydrolytic deamination of cytosines causes the formation of uracil (U), which basepairs with A rather than G (Fig: 1.4a). If an U:A mismatch

remains unrepaired before replication, it may manifest as a C:G to T:A transition. *In vivo* rates are difficult to estimate, because Cs in ssDNA deaminate with much higher rates compared to Cs in dsDNA, albeit it is thought to occur about 100-500 times per human cell per day (Frederico et al., 1990; Lindahl and Nyberg, 1974; Shen et al., 1994). The significance of this mutagenic process is highlighted by greatly increased C>T mutation frequencies in uracil-DNA glycosylase deficient cells (Sec. 1.4.2), which lack the ability to remove the altered base (Duncan and Weiss, 1982). In addition to spontaneous hydrolysis of cytosines, cells enzymatically induce the deamination of Cs in regular cellular processes such as somatic hypermutation (SHM) during antibody development due to activation-induced cytidine deaminase (AID) activity, and to mediate host defence against retroviruses by members of the apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) enzyme family.

Hydrolytic deamination of 5-methyl cytosines escape the base excision repair pathway

More problematic are hydrolytic reactions at 5-methylcytosines (5-meCs), which often occur in CpG-islands of eukaryotic genomes. Deamination of 5-meC transforms the base to a T rather than an U (Fig. 1.4b), thus precluding it from detection by uracil-DNA glycosylase (Sec. 1.4.2). If the mismatch repair (MMR) pathway (Sec. 1.4.3) fails to repair the resulting T:G mismatch prior to replication, such a site may manifest as T:A transition. Moreover, deamination rates of 5-meCs have been found to be threefold higher than that of unmethylated C, making the former especially vulnerable to this mutagenic process (Lindahl and Nyberg, 1974).

Hydrolytic cleavage of the N-glycosidic bond between the sugar and base components of a nucleotide produce abasic sites

Spontaneous hydrolysis often targets the base-sugar N-glycosidic bond leaving behind what is called an apyrimidinic or apurinic site (AP-site) (Lindahl et al. (1993), Fig. 1.4c). The total number of AP-sites generated in a single cell is estimated to be larger than 10,000 per day, and mostly due to depurinations which are 20× more rapidly released in comparison to pyrimidines. Unrepaired AP-sites may cause replication stalling with subsequent error prone translesion synthesis in which a low accuracy DNA polymerase inserts a random base opposite to an AP-site (Sec. 1.4.4).

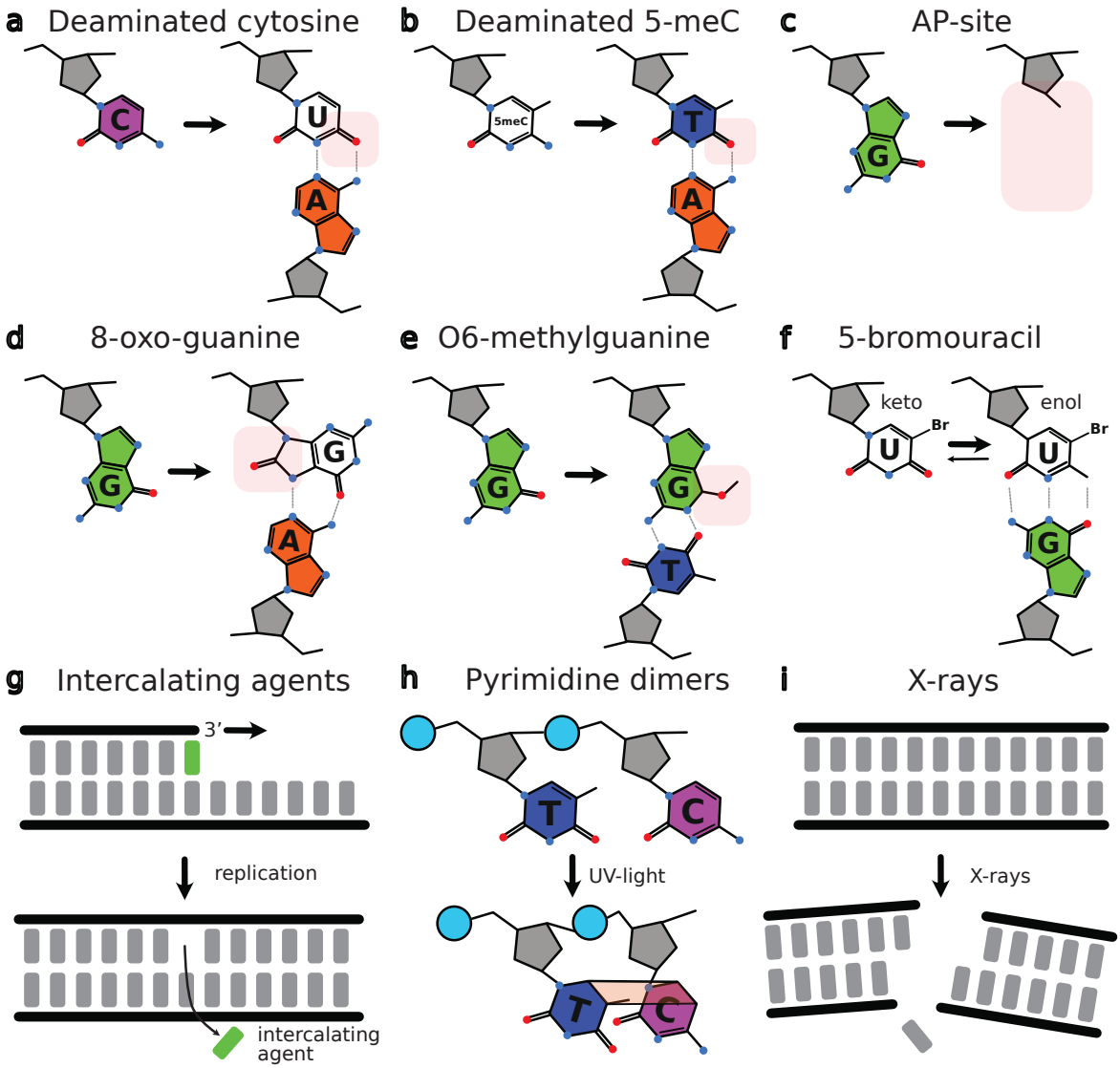


Fig. 1.4 **Endogenous and exogenous DNA lesions. a**. Hydrolytic deamination of cytosine turns the base into uracil which base pairs with adenine rather than guanine. **b**. The additional methyl group present in 5-methyl-cytosine turns the nucleobase into a thymine after deamination. **c**. Hydrolytic deamination may also occur at the N-glycosidic bond connecting the base to its ribose molecule, leaving behind an abasic-site. **d**. The oxidation of guanines may lead to the formation of 8-oxo-dG which base pairs with adenine rather than cytosine. **e**. Methylation of guanine produces O6-methylguanine, which may interact with thymine instead of cytosine. **f**. 5-bromouracil stabilises its enol enabling the base to predominantly base pair guanines. **g**. Intercalating agents may cause deletions. **h**. UV radiation introduces pyrimidine dimers by covalently linking two adjacent bases. **i**. X-rays may introduce double strand breaks.

Cellular byproducts engage in oxidativve reactions with DNA

Another source of endogenous DNA damage stems from reactive oxygen substances (ROS), which may be produced as a byproduct of metabolic reactions. ROS like superoxide (O_2^-),

hydrogen peroxide (H_2O_2), hydroxyl radicals (OH^-), and singlet oxygen ($^1O_2$) oxidise bases and may sometimes even cause single or double-strand breaks if cellular antioxidant defence mechanisms fail to neutralise these compounds (De Bont and Van Larebeke, 2004; Devasagayam et al., 2004).

It was estimated that 7,8-dihydro-8-oxoguanine (8-oxo-dG) is the most common modified base in human genomes, with steady state levels ranging from 0.07 to 145.25 adducts per Mbp (Higuchi and Linn, 1995; Spencer et al., 1996). The altered structure of 8-oxo-dG allows the compound to preferentially base pair with A rather then C (Fig. 1.4d), inducing C:G to T:A transversions when cells fall short to repair the alteration before the next round of replication.

1.3.2 Exogenous sources of DNA damage

Exogenous DNA damage occurs when chemical agents and environmental factors interact with DNA. Examples of chemical substances with genotoxic properties include polycyclic aromatic hydrocarbons (PAHs), alkylating agents, base analogs, and intercalators, whereas environmental factors comprise ionising radiation such as UV-light and X-rays. In contrast to endogenous DNA damage, which is usually confined to hydrolytic deamination and oxidation reactions, exogenous agents interact in various ways with DNA, and may introduce more severe types of DNA lesions such as double strand breaks (DSBs). Interestingly, the genotoxic effect of these substances often make these chemical agents useful in anti-cancer treatment. Here, I will briefly present the most common sources of exogenous damage, mention the utility of certain substances in context of cancer therapy, and point out the DNA repair mechanisms which help to restore induced DNA lesions.

Polycyclic aromatic hydrocarbons present in tobacco smoke turn into potent carcinogens after passing the liver

PAHs are widely distributed carcinogens found in tobacco smoke, car exhaust fumes, or charred food. Their structure usually contains two or more aromatic rings and are known to be generally inert (Baird et al., 2005; Boffetta et al., 1997; Boström et al., 2002; Kew, 2013). However, as soon as PAHs reach the liver, the P-450 system converts these molecules to reactive intermediates, capable of binding to DNA to create bulky adducts, which interfere with transcription or replication (Xue and Warshawsky, 2005). Prominent examples of PAHs include benzo(a)-pyrene, present in tobacco smoke, and aflatoxins which is a food contaminant (Boström et al., 2002).

Alkylating agents cause base modifications or DNA crosslinks

Another class of chemical species that may cause DNA damage are alkylating agents, which may origin from physiological processes such as DNA methylation (De Bont and Van Larebeke, 2004), but also happen to be anticancer treatment. In context of DNA damage, alkylating agents may be subdivided into mono and bifunctional agents, depending on the number of bonds they can form to nucleobases. Monofunctional methylating agents predominantly target nitrogen atoms (80 %), and only sometimes the oxygen atom of nucleobases. Cells repair adducts formed at N-atomes with the help of base-excision repair (BER) (Sec. 1.4.2), nucleotide-excision repair (NER) (Sec. 1.4.2) and alpha-ketoglutarate-dependent dioxygenase (AlkB) homologues. In contrast, bifunctional alkylating agents induce more complex and potentially deleterious inter-strand cross-links which require NER to be resolved.

For example, the highly mutagenic O^6-methylguanine (O^6-meG) mispairs T instead of a C, which may lead to G to A transitions (Fig. 1.4e). For this reason, cells have dedicated a single protein O^6-methylguanine-DNA methyltransferase (MGMT) (Sec. 1.4.1), whose sole purpose is to detect and remove O^6-meG from the genome (Kondo et al., 2010).

Base analogs mimic the structure of nucleotides allowing them to be incorporated into DNA by DNA polymerases or to interfere with nucleotide metabolism

Base analogs are derivatives of nucleobases with high structural similarity. This makes them useful in cancer therapy, as tumour cells have elevated rates of nucleotide turnover due to their propensity to divide faster in comparison to normal cells. This may lead, for example, to an intolerable accumulation of base analog-induced mutations in DNA, or to the inhibition of enzymes involved in the biogenesis of nucleotides (Tsesmetzis et al., 2018).

Prominent examples for base analogs include 5-bromouracil, which stabilises its enol rather than its keto form, enabling the uracil derivative to basepair with G instead of an A (Fig. 1.4f, Kaufman (1984)). 6-mercaptotpurine inhibits purine nucleotide synthesis and is used to fight acute leukemia (Karran and Attard, 2008). Finally, 5-fluorouracil is a thymine analogue which inhibits the biosynthesis of T, which is especially deleterious for fast dividing cells as the lack of dNTPs brings replication to a standstill (Miura et al., 2010).

Intercalating agents cause the formation of deletions or insertions

Intercalators are typically large and planar ring-shaped molecules, which do not damage DNA by breaking bonds or modifying bases, but harm the double helix by fitting in between adjacent base pairs. To incorporate intercalating agents, DNA must be partially unwounded to

open space between its bases. These conditions are met during replication where intercalators may cause deletions or insertions dependent on whether they sneak into the nascent or template strand DNA, respectively (Fig. 1.4g, Ferguson and Denny (2007); Wakelin (1986)). Moreover, intercalating agents may induce rearrangement events by interfering with the action of topoisomerases trying to unwind supercoiled DNA (Nitiss, 2009).

Intercalating agents such as etidium bromide, acridine orange or proflavin are used as nucleic acids stains, because their aromatic rings make them good chromophores. Some intercalators are used in chemotherapeutic treatment to slow down DNA replication in cancers such as Hodgkins' lymphoma (Mišković et al., 2013).

Ionising radiation causes double strand breaks and bulky DNA adducts

Ionising radiation is emission in form of traveling particles (neutrons, α or β-particles), or electromagnetic waves (higher ultraviolet (UV) spectrum, γ or X-rays) with sufficient energy to detach electrons from atoms or molecules (Borrego-Soto et al., 2015). In this way, ionising radiation may destabilise DNA directly by altering or breaking the structure of the double helix, or indirectly by turning inert molecules to reactive species such as ROS (Sec. 1.3.1) (Lomax et al., 2013).

Exposure to sunlight causes the formation of so-called UV-products at genomic sites where two pyrimidines are adjacent to each other. Mechanistically, UV-light triggers a "ring-opening" event allowing the double-bonds present in neighbouring pyrimidines to form covalent bonds with the residue right next to it (Fig. 1.4h). It is estimated that up to 50-100 mutagenic reactions per second occur in a single skin cell during UV exposure. Two common UV-light induced lesions are cyclobutan pyrimidine dimers (CPDs), mostly involving two thymidines, and 6-4 photoproducts (6-4PPs) which arise at sites where Cs and Ts are next to each other (Ikehata and Ono, 2011). Since both lesion types may cause replicative polymerases to stall and induce error-prone translesion synthesis, cells employ NER (Sec. 1.4.2) to fix such lesions before the next round of replication.

X- and γ-rays introduce DSB by directly breaking bonds in the DNA sugar-phosphate backbone (Fig. 1.4i), or indirectly by facilitating the formation of reactive oxygen species. Both X and γ-rays are used in low dosages for medical imaging purposes and in cancer radiotherapy (Little, 1993; Lomax et al., 2013).

1.4 DNA repair mechanisms

Cells evolved mechanisms capable of monitoring DNA, recognising damage and inducing an appropriate repair response to protect themselves against DNA damages. In this section,

I will outline the most important DNA damage repair pathways, including direct reversal DNA repair (Sec. 1.4.1), excision repair (Sec. 1.4.2), mismatch repair (MMR, Sec. 1.4.3), translesion synthesis (Sec. 1.4.4) and repair mechanisms to handle DSBs (Sec. 1.4.5).

1.4.1 Direct reversal DNA repair

Perhaps the simplest, but also the most specific form of DNA repair is direct reversal DNA repair. These repair mechanisms rely on a single enzyme, detecting and directly repairing specific types of lesions.

The most prominent example for this type of DNA repair is MGMT, which removes inappropriately placed methyl groups from the oxygen at the sixth position of guanine (O^6-meG, see also Sec. 1.3.2 and Fig. 1.4e). MGMT does not act as a true enzyme, because the protein can only complete the process once, which highlights the importance of clearing this lesion from the genome (Yarosh, 1985). Another example is photolyase, which completes direct reversal repair of UV-induced thymine dimers (Sec. 1.3.2, Fig. 1.4h). Note that human cells, as opposed to bacteria, fungi and some animals, do not express photolyase, but rather rely on NER (Sec. 1.4.2) to fix UV induced alterations (Li et al., 1993).

1.4.2 Excision repair

In contrast to direct reversal DNA repair, excision repair mechanisms are more versatile regarding the types of DNA lesion they recognise and repair. This makes them mechanistically more complicated, although the two main types of excision repair, BER and NER, can be conceptualised into three main phases: recognition, excision and repair. BER tends to repair endogenous DNA damages resulting from sources such as reactive oxygen species or depurination events. These alterations have in common that they do not cause large distortions to the DNA double helix. In contrast, NER focuses on helix distorting DNA modifications which are often due to exogenous genotoxins. Examples for such alterations are UV-induced pyrimidine dimers and bulky base adducts created by mutagens like aflatoxin B1 or polycyclic hydrocarbons.

Base excision repair

BER is initiated by DNA glycosylases, which recognise and remove a range of base modifications by cleaving off the glycosidic bond connecting the aberrant base to the DNA backbone. For example, DNA uracil glycosylase recognises Us that are usually not present in DNA, but may arise due to spontaneous deamination of cytosines (Sec. 1.3.1, Fig. 1.4a).

AP (apurinic/apyrimidinic) endonucleases subsequently cut the phosphodiester bond 5' to AP-sites that were left behind by DNA glycosylases. This nick serves as the substrate for weakly processive 5' to 3' exonucleases, which remove the AP phosphodiester backbone and roughly up to 3 bases. DNA polymerase β (Pol β) fills the single stranded gap in the DNA with complementary nucleotides, and DNA ligase restores the molecule by forming a phosphodiester bond between adjacent nucleotides (Braithwaite et al., 2005; Weinberg, 2006).

Nucleotide excision repair

There are two activation pathways which trigger DNA damage repair by NER in a genome location dependent manner. While global genome NER (GG-NER) surveils the entire genome including coding and non-coding regions, a second mechanism termed transcription coupled NER (TC-NER) provides additional protection for transcribed regions and is triggered by stalling of RNA Pol II (Schärer, 2013).

A multitude of different proteins including XPC+DDBI/XPE are involved in initiating GG-NER in eukaryotes. This protein complex binds to the strand opposite of the lesion, mediates DNA opening by recruiting TFIIH (Lee et al., 2014), and attracts various other factors such as RPA or XPA. Eukaryotes employ two distinct endonucleases to excise a 24-32 bp long stretch of ssDNA including the lesion. XPA recruits ERCC1-XPF to make the first incision on the 5' side of the bubble, while XPG mediates the second cut on the respective 3' side (Fagbemi et al., 2011). The excised oligonucleotide is subsequently removed by DNA helicase activity of TFIIH. In the final step of NER, replicative Pol ε or Pol δ fill the excised gap (Lehmann, 2011; Ogi et al., 2010), and DNA ligases seal the nick between newly synthesised and adjacent nucleotides (Petruseva et al., 2014).

TC-NER couples NER to transcriptional activity (Sec. 1.2.2). In TC-NER, stalling of RNA Pol II triggers recognition of a DNA lesion, and recruits transcription coupled repair factor (TCRF), which uses ATP hydrolysis to displace RNA Pol II and subsequently recruits NER proteins (Schärer, 2013).

1.4.3 Mismatch repair

Like excision repair mechanisms, MMR follows the theme of orchestrating different proteins to conduct a three step process involving recognition of DNA damage, excision and subsequent repair. However, one challenge MMR has to overcome is to choose the parental strand before excising the mis-incorporated base. This is difficult, because in contrast to other DNA

assaults which come along with structural distortions, mismatch errors fail to display any obvious lesion marks (Heller and Marians, 2006; Pluciennik et al., 2010).

Mismatch recognition by MutS In eukaryotes, DNA damage recognition is mediated by Mutator S protein (MutS), which is a heterodimer composed of the subunits MutS homolog 2 (Msh2) and MutS homolog 6 (Msh6). MutS:ADP surveils the entire genome by scanning roughly 700 base pairs in a non-directional and energy independent manner. In ADP bound form, MutS temporarily binds to DNA with a half-life $\sim$1 s after which it will fall off. To detect mismatches and smaller indels, MutS encircles double stranded DNA while slightly bending the molecule. It is thought that mismatches and smaller indels allow the double helix to be bend more easily in comparison to intact DNA that tries to stay planar. Upon damage recognition, MutS:ADP exchanges ADP for ATP, which results in a conformational change allowing the protein to form a more stable complex with the selected DNA, thus marking a mismatch site (Fishel and Lee, 2016; Hsieh and Yamane, 2008).

Nick-directed nascent strand detection The next step of MMR deals with excising the wrongly inserted base in the newly synthesised DNA strand. A small number of bacterial species, including *E. coli* recognise nascent DNA by the temporal lack of DNA methylation. However, in most bacterial and eukaryotic cells, the newly synthesised DNA is identified by the presence of nicks (e.g. at Okazaki fragments) which facilitate sliding clamp loading. The orientation of proliferating cell nuclear antigen (PCNA) guides MMR endonucleases to cut the newly synthesised strand (Flores-Rozas et al., 2000; Iyer et al., 2008; Pluciennik et al., 2010).

Mismatch bound MutS:ATP recruits and activates Mutator L protein (MutL) by triggering its ATPase domain that hydrolyses bound ATP to ADP. Unlike the ATP-bound form, MutL:ADP exhibits an unspecific endonuclease activity. MutL and MutS then bind to PCNA which instructs MutL to cut newly synthesised DNA randomly close to the lesion. Quickly after MutL placed the cut, its bound ADP is exchanged with ATP thereby inactivating the enzyme (Kadyrov et al., 2006).

Nick-directed strand repair To remove the DNA containing the misincorporated base, the 5' to 3' exonuclease ExoI removes all DNA between the closest nick and the cut site of MutL. It is not very well understood how ExoI is recruited to these sites since the closest nick might often be not proximal to the mismatch site. However, as soon the strand is removed, it can readily be resynthesised by Pol δ and PCNA (Iyer et al., 2006).

1.4.4 Translesion synthesis

Replicative DNA polymerases easily stall at sites where they encounter corrupted bases in the template (Sec. 1.2.3). At this point, it is too late to excise the DNA damage, because this would lead to a break in the unwound DNA, and eventually cause a replication fork collapse with deleterious consequences. To ensure survival, cells employ translesion polymerases, which replicate through the damage by "guessing" what base to insert opposite to an aberrant base, thus having error rates up to 1 mis-incorporated base per 100 bases replicated (Knobel and Marti, 2011; McCulloch and Kunkel, 2008).

A commonality amongst translesion polymerases is the lack of a 3' to 5' exonuclease activity and a less restrictive active site capable of accommodating damaged DNA, e.g. due to the lack of an O-helix (Sec. 1.2.3, Knobel and Marti (2011)). Although many translesion polymerases are interchangeable, they replicate over different types of lesions with different efficiencies, and hence there are at least nine distinct translesion polymerases in mammalian cells (Knobel and Marti, 2011). Some polymerases are specialised to fill bases opposite to AP-sites, while others extend nascent DNA opposite to bulky DNA adducts. For example, pyrimidine dimers are preferably recognised by DNA polymerase η (Pol η), which by default places two As opposite to such lesions (Hendel et al., 2008; McCulloch et al., 2004; Takata et al., 2006). DNA polymerase κ (Pol κ) synthesises past 8-oxo-dG and tends to incorporate As more often than Cs (Irimia et al., 2009; Weinberg, 2006).

Polymerase switching

Stalled replication forks initiate translesion synthesis by switching PCNA bound Pol ε or Pol δ with a translesion polymerase. In comparison to replicative DNA polymerases, error-prone polymerases are less processive, causing them to fall off the sliding clamp shortly after synthesising across the lesion (Lovett, 2007).

1.4.5 Double strand break repair

Double strand break repair (DSBR) is essential, as DSBs are the most toxic form of DNA damage. The two major pathways for DSBR are non-homologous end joining (NHEJ) and recombinational repair (RR). Under normal circumstances, cells deal with most of the double strand breaks by employing the *classical* NHEJ pathway or RR. The choice between NHEJ and recombinational repair depends on the cell cycle phase of the cell. While NHEJ usually repairs most of the DSBs at any point in the cell cycle phase, RR takes over during S-phase, when the homologous DNA of the newly replicated chromosome is available (Scully et al., 2019).

Non-homologous end joining

Conceptually, the easiest way to fix a DSB is to directly concatenate both broken ends back together. To accomplish this, eukaryotic cells have two NHEJ pathways with different degrees of fidelity at their disposal.

Classical NHEJ Classical NHEJ is initiated by the heterodimeric surveillance protein Ku 70-90, which monitors DNA for DSBs and binds to double stranded ends (Britton et al., 2013). From there it recruits DNA protein kinases helping to keep both DSB ends together (Gottlieb and Jackson, 1993). DNA protein kinases also recruit and activate Artemis, a nuclease that processes the DNA ends. In the final step of classical NHEJ, DNA ligases seal the DNA (Ahnesorg et al., 2006; Nick McElhinny et al., 2000). Artemis activity in classical NHEJ may leave small scars on the sequence by adding or removing a few nucleotides (Xie et al., 2009). Although this may be harmful in protein coding parts of the genome, it has little impact in non-coding regions. For this reason, the classical pathway is considered to be high-fidelity (Fig. 1.5a, Scully et al. (2019)).

Alternative/Error prone NHEJ In contrast, alternative NHEJ is a low fidelity DNA repair mechanism which is often associated with chromosome translocation or loss. Cells employ this pathway in a final attempt to fix DSBs, for example when components of the classical pathway are missing or their recruitment is delayed (Scully et al., 2019).

PARP1 initiates alternative NHEJ by binding to the broken DNA ends and recruiting several repair proteins to polish bound DSB ends. However, rather than simply stitching the broken DNA ends back together, alternative NHEJ relies on microhomolgy, i.e. that the ends of concatenated DNA strings embed sequences that base pair to each other. This may pull together DNA ends from different chromosomal locations, which often breaks the integrity of large genomic segments (Fig. 1.5b, Lieber (2008)).

Recombinational repair

In comparison to NHEJ, RR is a complex process to resolve DSBs which subsumes multiple subpathways, each with specific enzymatic requirements. However, RR can be subdivided into three steps: presynapsis including DSB formation and DNA strand resection, synapsis encompassing strand invasion, and postsynapsis which can be further split into three different modes of homologous recombination (HR): break induced annealing (BIR), synthesis-dependent strand annealing (SDSA), and HR resolution via double Holliday junction (dHJ). RR of DSBs requires the presence of nearly identical dsDNA, which is provided by the intact

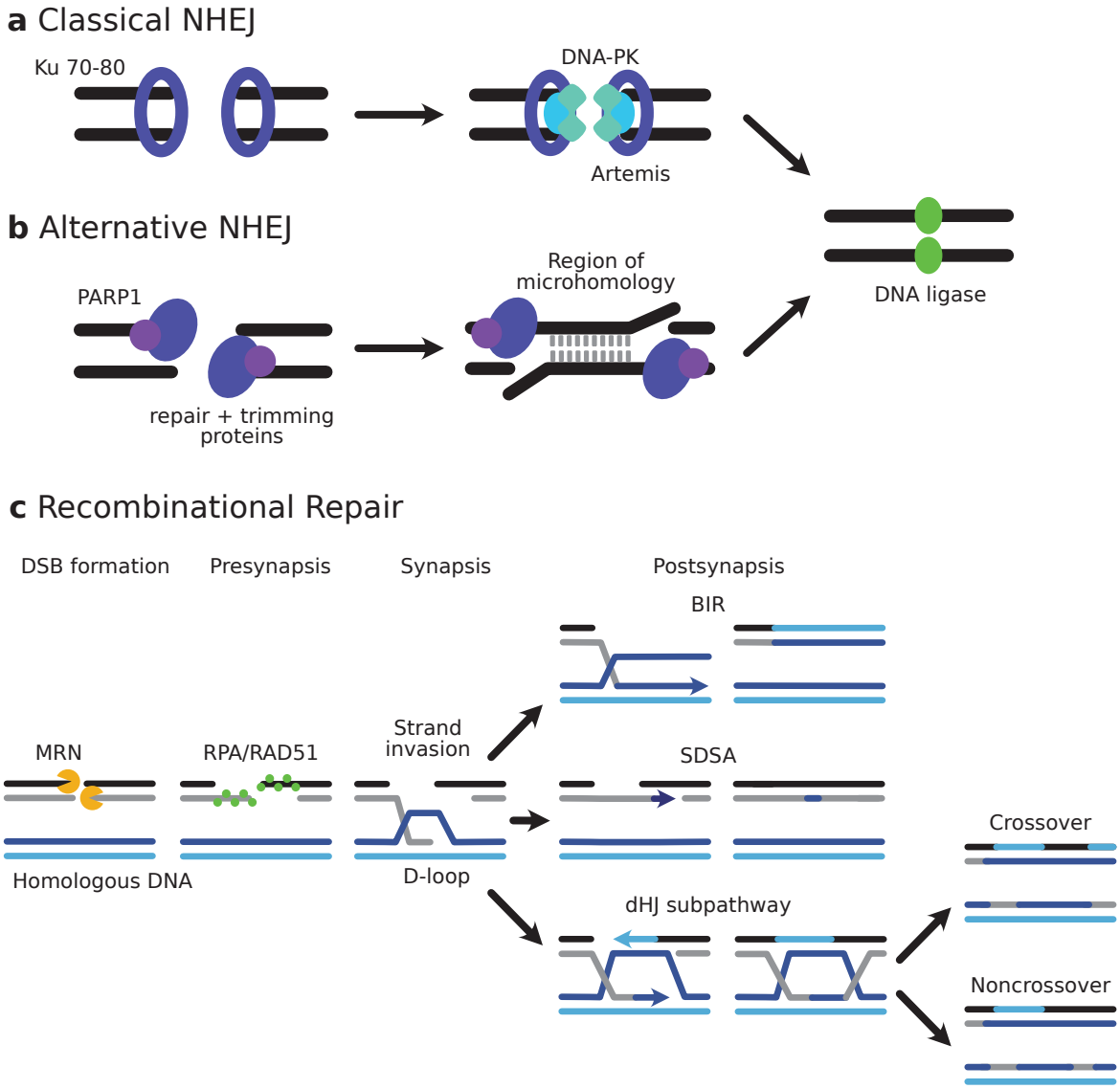


Fig. 1.5 **DNA double strand repair via NHEJ and recombinational repair. a**. Classical NHEJ recognises DSBs with Ku70-80 proteins, fixes broken ends with DNA protein kinases and Artemis and ligates broken ends with DNA ligases. **b**. Alternative NHEJ uses PARP1 and a few repair and trimming proteins to prepare DNA ends, potentially from different chromosomes, and joins them subsequently by forming a region of microhomology. **c**. Recombinational repair involves three major steps: presynapsis, synapsis and postsynapsis.

chromosome from either paternal or maternal origin, thus contributing possibly different alleles to the genomic region to be repaired (Heyer et al., 2010).

Presynapsis The first step in presynapsis is to resect the DNA at both sides of a DSB with a 5' to 3' exonuclease (MRN complex in humans), which results in long and single stranded 3'

overhangs. RPA's high affinity for ssDNA quickly attracts the protein to resected overhangs, eliminating secondary structures in ssDNA, and enabling RAD51 filament formation, which catalyses homology search and DNA strand invasion (Mimitou and Symington, 2009). However, RPA binding to ssDNA also displays a kinetic barrier, thus making so-called mediator proteins mandatory to ensure RAD51 assembly, amongst them BRCA2, the human breast and ovarian tumor suppressor protein (Fig. 1.5c, Heyer et al. (2010); Yang et al. (2005)).

Synapsis During synapsis, the RAD51 coated single stranded 3' overhang aligns to the homologous DNA of its sister chromatid, where the intermediate invades its complementary DNA to form Watson-Crick base pairs, thus displacing the originally aligned DNA strand to form a so-called D-loop (Fig. 1.5c, Heyer et al. (2010)).

Postsynapsis In postsynapsis, HR can progress in three different ways. In BIR the D-loop may be converted to a fully working replication fork if the other end of the DSB is missing. Although this pathway recovers the lost chromosome completely, BIR results in loss-of-heterozygosity since all alleles distal to the DSB are lost (Heyer et al., 2010). In SDSA the extended D-loop re-anneals to the other resected end of its original chromosome, thus avoiding crossovers and genomic rearrangements (Heyer et al., 2010; Pâques and Haber, 1999). Finally, via dHJ formation, which is rather rarely used in RR of somatic cells and was originally discovered in meiotic recombination. Here, directly after D-loop formation, the displaced DNA strand from the intact chromatid instead basepairs with the 3' overhang from the other side of the DSB. In this configuration, the 3' overhangs are elongated by DNA polymerases which use the DNA strands of the intact chromosome as a template to refill the depleted DNA. In the next stage, the branches migrate, thereby giving rise to the dHJ structure. Finally, the double Holliday junction is resolved by cleaving the DNA, which yields different recombinant products (patched or crossover) dependent on how the DNA was cut (Fig. 1.5c, Bzymek et al. (2010)).

1.5 Mutagenesis

The previous sections set the framework for the upcoming discussion on mutagenesis in genomic context. Mutagenesis results from a combination of DNA damage and failed or error prone attempts of DNA repair, and results in lasting mutations which may be detected in sequencing experiments as SNVs, multi-nucleotide variants (MNVs), indels or structural variants (SVs). Importantly, sequencing studies revealed that mutation rates vary across

the human genome. In fact, mutagenesis is affected by various factors ranging from DNA organisation to epigenetics, and fundamental cellular processes such as transcription and replication.

1.5.1 DNA organisation affects mutagenesis

The nucleosomal organisation of DNA protects the molecule against spontaneous mutations, helps to regulate oncogenic genomic regions, and affects intra nucleosomal mutation rates.

Nucleosomal packaging surpresses spontaneous hydrolytic deamination of cytosines

Nucleosomal-bound DNA undergoes fewer C>T mutations, because of suppressed hydrolytic deamination of cytosines (Sec. 1.3.1), and exhibits lowered rates of G>T and A>T mutations relative to nucleosome-depleted DNA (Chen et al., 2012). These effects are reflected by fewer SNVs at highly positioned nucleosomes (Schuster-Böckler and Lehner, 2012).

Aberrant expression of H1 linker histones leads to the expression of transcriptional programs that favour cencer cell renewal

Linker histone induced higher order DNA structures protect the integrity of the genome, as it has been shown that aberrant expression of H1 has links to cancer. Particularly, absence of the H1 histone leads to decreased nucleosomal DNA interactions in AT-rich genomic regions, which is coupled to the derepression of close genes, resulting in expression of transcriptional programs that support cancer cell renewal (Torres et al., 2016).

The mode of nucleosomal DNA-binding affects mutagenesis

Pich et al. (2018) found that tumours like oesophageal adenocarcinomas, gastric cancers and malignant lymphomas show a relative increase of mutation rates peaked at minor grooves facing towards histones (AT-rich), while cancers like skin melanoma and lung adenocarcinoma show the opposite pattern, with maxima of mutation rates in the minor groove facing away from the nucleosome (GC-rich, Fig. 1.1). The analysis of mutation types found at nucleosome-bound minor grooves revealed that mutation enrichment in aforementioned cancers could be linked to mutation types that are characteristic for the respective tumour. For example, oesophageal cancers exhibit elevated rates of T>C mutations in minor grooves facing towards nucleosomes, while C>A and C>T base substitutions, characteristic for lung cancers and skin melanomas, tend to occur with an unexpected high rate at minor grooves facing away from the histone octamer.

1.5.2 Epigenetic modulation affects mutagenesis

Epigenetic states enable cells to modulate genomic activity flexibly, and are yet persistent through multiple cell divisions. For this reason, epigenetic states have a great influence on mutagenesis, for example, by providing cancers the means to modify cellular properties by gaining epigenetic control of the genome to express oncogenic traits.

Histone marks correlate with mutation rates

Schuster-Böckler and Lehner (2012) showed that histone marks and other genomic features are highly correlated with local SNV densities. Marks associated with active chromatin, particularly H3K4me3 and H3K9ac, anti-correlate with mutation rate densities, while repressive chromatin due to H3K9me3, H3K9me2 and H4K20me3 correlate positively with base substitution rates. They also revealed that a combination of genomic features may even explain up to 60 % of the variance in cancer mutation rates, indicating interdependencies among different chromosome features (Schuster-Böckler and Lehner, 2012). Additionally, Polak et al. (2015) showed that chromatin states accounting for the cell type of the corresponding cancer, as well as replication timing, explain even up to 86 % of the variance in mutation rates across human genomes.

Mutagenesis and DNA methylation

DNA methylation may influence mutagenesis in cancer genomes in various ways. On the one hand, 5-meC undergoes spontaneous deamination at higher rates than its unmethylated counterpart, which inevitably leads to C>T substitutions if MMR fails to repair the resulting mismatch (Sec. 1.3.1). On the other hand, aberrant DNA methylation may influence mutagenesis indirectly by silencing tumour suppressive genes such as the *MLH1* gene involved in MMR, thus increasing mutation rates by several orders of magnitude in MMR-deficient cells[5]. Moreover, severe hypomethylation seems to be associated with genomic instability, as DNA methyltransferase deficiency in mouse models display higher frequencies of structural mutability (De and Michor, 2011; Li et al., 2012), and germline mutations in the DNA methyltransferase *DNM3B* gene cause the ICF syndrome, characterised by centromeric instability (Okano et al., 1999).

[5]Epigenetic silencing of DNA repair is not limited to *MLH1*, but has also been reported for several other genes including *MGMT*, *BRCA1*, *WRN*, *FANCF* and *CHFR* (Toyota and Suzuki, 2010).

1.5.3 Mutagenesis in context of transcription

Actively transcribed genes and associated regulatory regions are especially important to cells as disruptive mutations in coding regions may have deleterious consequences. For this reason, transcription encourages the activity of DNA repair mechanisms such NER and MMR.

DNase I hypersensitivity assays reveal hypo-mutated promotor regions

Open chromatin regulatory regions, including promotors, enhancers and insulators, have been extensively mapped in diverse organisms using DNase I hypersensitivity seq assays (Thurman et al., 2012), and been associated with reduced local densities of somatic mutations in a range of cancer genomes. The reduction of SNV frequencies in DNase I hypersensitive sites (DHSs) compared to the genome average could not be attributed to confounding factors such as differences between intergenic regions and genes, the distance to TSS, replication timing, GC content or distances to telomeres or centromers, which suggested excision repair mechanisms as the most likely explanation for the local mutation rate reduction (Polak et al., 2014). Moreover, the accessibility of DHSs encourages the activity of DNA repair pathways such as NER or BER, as both mechanisms have been reported to prefer free DNA over nucleosomal bound DNA (Amouroux et al., 2010; Bell et al., 2011). The analysis of four NER deficient melanoma samples revealed increased mutation rates at DHSs, thus supporting the theory that GG-NER is responsible for the hypomutation phenomenon at gene regulatory regions (Polak et al., 2014).

RNA Pol II and transcription factor binding increases local mutation rates within promotors

Interestingly, binding of TF to transcription factor binding sites (TFBSs) leads to $5\times$ larger mutation rates at TFBS in comparison to neighbouring flanking DNA in skin melanoma, thereby refining the hypomutation phenomenon observed in DNAase hypersensitive genomic regions. Local elevation in mutation rates do not affect inactive TFBS, is present at distal TFBS (enhancers), and was also detected in a normal skin sample, indicating that mutagenesis due to protein occupancy is a normal process, and not necessarily bound to the deregulated environment of a cancer cell. Further investigations in UV-irradiated cell lines subjected to XR-seq, which captures NER excised DNA fragments, revealed that elevated mutation rates observed at active TFBS are caused by a decrease of local NER activity (Sabarinathan et al., 2016).

Transcription-coupled nucleotide excision repair induces asymmetric mutagenesis

Imbalances in the distribution of mutations on template and coding strand DNA are mostly caused by TC-NER (Sec. 1.4.2), which repairs DNA damages at sites where RNA Pol II stalls (Donahue et al., 1994; Hanawalt and Spivak, 2008). Also, transcription levels anti correlate with mutation rates on template strand DNA, consistent with the notion that increased rounds of transcription provide more opportunity to clear lesions (Chapman et al., 2011; Lawrence et al., 2013; Pleasance et al., 2010a). Conversely, coding strand DNA may be more susceptible to mutations as it remains single-stranded during transcription, and lacks the surveillance of TC-NER (Jinks-Robertson and Bhagwat, 2014). Transcriptional asymmetries are particularly noticeable in tumours with high exposure to exogenous agents such as UV-radiation or smoking, but are also evident in liver cancers (Haradhvala et al., 2016).

DNA mismatch repair decreases mutation rates in exons

Frigola et al. (2017) found fewer mutations in exonic genomic regions in seven tumour types, and especially in genomes with mutations in the *POLE* gene. They also observed the colocalisation of the histone modification H3K36me3 in exonic regions, which has been implicated in the recruitment of MutS of the MMR pathway. For this reason, the authors speculated that MMR is responsible for lowered mutation rates in exons, and verified this hypothesis by assessing exonic mutation rates in microsatellite instability (MSI) tumours lacking MMR, which indeed showed elevated SNV frequencies in exons.

1.5.4 Mutagenesis and DNA replication

Replicative DNA polymerases in conjunction with MMR suppress error rates to 1 mistake per 10^{10} replicated base pairs (Kunkel and Bebenek, 2000). However, there are multiple factors ranging from local nucleotide composition to defective DNA polymerases that may cause substantial variation in mutation rates due to replication.

Repetitive DNA sequences are prone to replication errors

Eukaryotic genomes harbour a great abundance of repetitive DNA elements, termed microsatellite sequences, which tend to favour mispairing during replication. This results in higher mutation rates at microsatellites in comparison to genomic regions with a more complex sequence composition (Jeffreys et al., 1988).

Mutations in the exonuclease domain of DNA polymerase epsilon increase mutation rates significantly

Mutations in the $POLE^{exo}$ domain may have dramatic effects on genomes and cause a so-called hypermutator phenotype. The most frequent mutations are P286R (P=proline, R=arginine), possibly leading to polymerase hyperactivity and increased tolerance to mismatches due to a decreased affinity of the exonuclease domain to ssDNA, and V411L (V=valine, L=lysine) which functional implications are spurious, as this residue lies far away from the active site and does not directly interact with DNA (Parkash et al., 2019; Xing et al., 2019).

Some mutational processes introduce mutations with replicational asymmetries

Imbalances in the distribution of mutations on leading and lagging strand DNA may result from differential processing of both parental strands by distinct enzymes, or due to the extended time frame lagging strand DNA has to endure in vulnerable ssDNA form. Haradhvala et al. (2016) found replicational asymmetries in tumours with *POLE* mutations, MSI and APOBEC associated tumours. Replicational asymmetries in these tumours were most pronounced in early-replicating regions, and strand-specific mutation rates flipped sign at replication timing minima (i.e. ORIs).

Replication timing affects mutagenesis

Studies investigating the association of replication timing with the evolutionary divergence and SNP densities in human populations found that late replicating regions accumulated more substitutions over the course of evolution (Chen et al., 2010; Stamatoyannopoulos et al., 2009; Wolfe et al., 1989). The analysis of high-resolution maps of human replication timing programs underpinned these findings by revealing strong correlations between base substitution rates and replication timing, and additionally found that structural variation tends to accumulate in early replicating regions (Koren et al., 2012). Supek and Lehner (2015) identified variable MMR as the basis for the changes of base substitution rates in early and late replicating regions by analysing mutation rates at megabase scale. This analysis revealed that mutation rates are largely stable across cancer types and replication timing states, but start to deviate in samples with inactivated MMR, in which late replicating regions are no longer enriched in mutations.

Inclusion of ribonucleotides during DNA replication

Ribonucleotides exhibit high structural similarity to dNTPs, as they only differ by a single hydroxyl group on the 2' carbon of their sugar component. Failure of DNA polymerases to accurately discriminate between both compounds often leads to the mis-incorporation of ribonucleotide triphosphates (rNTPs), which is not limited to, but most often due to much greater nuclear levels of rNTPs relative to dNTPs (500-3000 μM vs. 12-30 μM). While occasionally embedded ribonucleotide monophosphates (rNMPs) do not introduce severe DNA distortions, their presence may impede progression of the replication machinery, interfere with nucleosome assembly, or alter protein binding to DNA. For this reason, cells remove incorrectly incorporated rNMPs via the proofreading function of DNA polymerases and BER (Cerritelli and Crouch, 2016).

1.6 Mutational signature analysis

Although only a small number of mutations confer a selective advantage, i.e. represent mutations that promote oncogenesis, the totality of mutations is characteristic for the sum of the mutational processes that were acting in the cancerous cell. The recent advancement in next-generation sequencing gives access to this information and hence there is a growing interest to utilise this data to better understand the nature and genomic impact of various mutational processes.

1.6.1 Mutational Signature Analysis

A particularly successful method pioneered by Alexandrov et al. (2013a) is mutational signature analysis, which extracts common characteristic base substitution patterns from a set of cancer genomes. Several basic types of base changes (C>A, C>G, C>T, T>A, T>C, and T>G[6]) were long known to be characteristic for certain tumour types. For example, UV-light induced pyrimidine dimers often manifest as C:G to T:A (C>T) mutations in skin melanoma, whereas lung cancers suffer from high loads of G:C to T:A (C>A) substitutions due to the carcinogens present in tobacco smoke. Since the formation of base substitutions is strongly affected by the immediate sequence context, it is sensible to incorporate the flanking 5' and 3' base to the classification, which gives rise to a 96 trinucleotide classification ($4\times\ 6\ \times\ 4$). In mutational signature analysis, the count of each trinucleotide mutation type and sample is expressed in form of a matrix and then subjected to NMF, a mathematical procedure

[6]To avoid ambiguity due to the symmetry of nucleobases, mutations are always specified with respect to the involved pyrimidine.

which extracts a set of prototypical mutation patterns that are common to a large fraction of cancer genomes, and their respective mutation loads in each of the samples. Mathematically, this corresponds to a matrix factorisation, in which a catalogue of cancer genomes **C** is decomposed into a set of mutational signatures **S** and their constituent activities or exposures **E**

$$\mathbb{E}[\mathbf{C}] = \mathbf{S} \times \mathbf{E} \quad \text{where } \mathbf{C} \in \mathbb{N}_0^{p \times n}, \mathbf{S} \in \mathbb{R}_+^{p \times s}, \text{and } \mathbf{E} \in \mathbb{R}_+^{s \times n} \tag{1.1}$$

where p is the number of mutation types (usually $p = 96$), n the number of cancer genomes and s the number of mutational signatures.

1.6.2 Characterised mutational signatures

One reason for the great success of mutational signature analysis is the high interpretability of identified signatures. The following section reviews the most important SBS signatures of the Catalogue of Somatic Mutations in Cancer (COSMIC) catalogue.

SBS1 reflects a deamination signature with clock-like properties

One of the most pervasive phenomena in eukaryotic genomes, the deamination of 5-meC (see Sec. 1.3.1), is captured by COSMIC SBS signature 1, which depicts a highly characteristic N[C>T]G pattern indicative for C:G to T:A transitions at NpCpG (Alexandrov et al., 2013b). SBS1 has been reported in at least 25 different cancer types and has clock-like properties, i.e. the mutation load due to SBS1 is proportional to the age of the patient, indicating that deamination of 5-meC is independent from cancer driving oncogenic alterations or certain exogenous mutagenic sources, but rather adds constantly a certain number of mutations per year (Alexandrov et al., 2015). Interestingly, SBS1 mutation rates differ in cancers, as epithelial tissues with high turnover exhibit largest rates, making SBS1 mutations indicative for a "molecular clock" that registers the number of mitoses cells experienced, as replication with failure of MMR is required to manifest the mutational pattern (Alexandrov et al., 2015).

APOBEC mutagenesis (SBS2 and SBS13)

Another mutational process related to the deamination of cytosines is enzymatically catalysed by members of the APOBEC enzyme family, whose physiological responsibilities include the restriction of retroviridae and mobile retroelements. APOBEC enzymes recognise ssDNA and mutate cytosines in TpC sequence context, with strong preference for A or T 3' flanking bases, leading to the characteristic T[C>K]W (W=A/T, and K=G/T) pattern of SBS2/13

(Taylor et al., 2013). This mutational process was first characterised in breast cancers (Nik-Zainal et al., 2012), but was also found in 16 other cancer types (Alexandrov et al., 2013a). It was shown that T[C>K]W mutations exhibit a high degree of strand coordination, indicative for APOBEC mutagenesis on lagging strand DNA during replication. Another property of APOBEC mutagenesis is the tendency to generate highly clustered mutations, a phenomenon termed kataegis[7], which often colocalises with rearrangement breakpoints, suggesting that the enzyme readily processes ssDNA at the sites of DSB.

The analysis of the higher order nucleotide context of APOBEC characteristic base substitutions revealed a prevalent YT[C>K]A motif, indicative for the APOBEC3A isoform[8] (Chan et al., 2015). Consistent with APOBEC3A as the main contributor to mutagenesis, it was found that individuals carrying at least one copy of an APOBEC3B deletion polymorphism have a 2.37-fold increased relative risk for cancers with large contributions of SBS2/13 (Nik-Zainal et al., 2014). Finally, although SBS2 and SBS13 are most likely due to DNA damage by APOBEC, the representation in two signatures with predominant contribution of C>T and C>G mutations respectively, may be caused by differences in the involvement of DNA repair mechanisms and translesion polymerases.

SBS3 is associated with homologous recombination deficiency

SBS3 exhibits an almost equal representation of all 96 mutation types, and is strongly associated with *BRCA1* and *BRCA2* mutations, which play important roles in the mediation of HR (Sec. 1.4.5). However, SBS3 may also contribute substantial amounts of mutations in samples which do not exhibit any deficiencies in the aforementioned genes, indicating that other mechanisms independent from *BRCA1/2* may generate it (Alexandrov et al., 2013b). Particularly, germline nonsense and frameshift mutations in *PALB2*, which binds and localises BRCA2 to the nucleus, and silencing of *RAD51C* and *BRCA1* via promotor methylation were shown to induce SBS3 mutations as well (Polak et al., 2017).

Tobacco smoking manifests as genomic C>A mutations (SBS4)

SBS4 is characterised by C>A mutations and is most likely due to tobacco smoking. This signature was initially discovered in lung adeno, squamous and small carcinomas, head and neck, and liver cancers. The signature exhibits a transcriptional strand bias, indicating higher mutation probabilities for C>A mutations on transcribed DNA, which is consistent with the propensity of many tobacco carcinogens to form adducts with guanine (Alexandrov et al.,

[7] Kataegis is the greek word for thunderstorm (Nik-Zainal et al., 2012).
[8] As opposed to the APOBEC3B specific motif RT[C>K]A.

2013b). Moreover, SBS4 shows strong similarities to the mutational spectrum observed in cells exposed to polycyclic aromatic hydrocarbons (Alexandrov et al., 2016).

Deficiencies in the NER pathway may induce a flat mutational spectrum (SBS5)

SBS5 displays a relatively flat spectrum with a slight predominance of C>T and T>C mutations. The signature has been proposed to be associated with tobacco carcinogens, but was also found in cancers with no relation to tobacco consumption, thus making the concrete aetiology of the signature unclear (Alexandrov et al., 2013b). SBS5 exhibits clock-like properties in some cancers, thus contributing a constant rate of mutations, which range from 31.8 mutations/GB/year in papillary cell kidney caner to 2.8 in acute myeloid leukemias (Alexandrov et al., 2015). Kim et al. (2016) proposed an association of SBS5 to NER (Sec. 1.4.2), because a signature enrichment analysis revealed an association to the *ERCC2* gene, which encodes a helicase that is needed to unwind DNA adjacent to a bulky DNA lesion.

Various mutational signatures indicate mismatch repair deficiency (SBS6/14/15/16/20/21/26/44)

A selection of mutational signatures are associated with the co-occurrence of small deletions at nucleotide repeats. This phenomenon is known as microsatellite instability and is characteristic for tumours with mismatch repair deficiency (MMRD) (Sec. 1.4.3). SBS6, for example, is characterised by predominantly C>T at NpCpG, and its presence in colorectal cancers is strongly associated with the inactivation of MMR genes. A very similar signature, SBS15, with a greater prominence of C>T at GpCpN sites is found in lung and stomach cancers (Alexandrov et al., 2013b). Moreover, Haradhvala et al. (2018) suggested the concurrent loss of DNA polymerase proofreading proficiency and MMR as aetiology of SBS14 and SBS20.

UV-light induced mutational signatures (SBS7a/b/c/d)

SBS7 reflects mutagenesis due to UV-light, and was first described in malignant melanoma with a strong transcriptional strand bias, consistent with the notion that UV-light induces bulky adducts which are repaired with TC-NER (Alexandrov et al., 2013b). Hayward et al. (2017) identified two signatures with high similarity to SBS7, thus suggesting a strong link to UV-light exposure, possibly representing different photoproducts: SBS7a consists predominantly of C>T at TCN, while SBS7b has larger contributions of C>T mutations in CCN context. The latest update of the COSMIC catalogue of mutational signatures introduced two novel UV-light signatures termed SBS7c and SBS7d. These signatures are characterised by T>A at NTT and T>C at NTT mutations, respectively (Alexandrov et al., 2020).

Mutations in the exonuclease domain of the leading strand polymerase trigger SBS10

SBS10 produces large numbers of base substitutions in colorectal and uterine cancers, and has been associated with mutations in the exonuclease domain of Pol ε (Alexandrov et al., 2013b). The latest instalment of the COSMIC catalogue splits this signature into two spectra SBS10a and SBS10b, which resemble the spectral C>A and C>T parts of SBS10, respectively (Alexandrov et al., 2020).

Mutational signatures associated with cancer treatments (SBS11 and SBS17a/b)

Temozolomide is an alkylating agent that is used to treat melanoma and tumours of the central nervous system. The drug is quickly absorbed and converted to an active compound that forms several DNA adducts, which are usually detected and cleared by BER or MGMT (Poon et al., 2014). SBS11 displays a characteristic N[C>T]C pattern which is likely to represent the effects of temozolomide treatment, as the signature was discovered in pretreated malignant melanoma and glioblastoma multiforme (Alexandrov et al., 2013b).

Fluoropyrimidines are one of the most commonly used anticancer treatments of solid cancers, which inhibit the growth of tumours by interfering with their nucleotide metabolism. Christensen et al. (2019) identified fluorouracil (5-FU) treatment as the most likely cause for the mutational pattern of SBS17, which is characterised by a striking peak of T>G mutations in a CTT trinucleotide context. The latest version of the COSMIC catalogue tears apart the T>C and T>G part of SBS17 into two signatures SBS17a and SBS17b (Alexandrov et al., 2020).

Alcohol consumption may introduce characteristic T>C mutations at ApT dinucleotides in liver cancers (SBS16)

SBS16 depicts a very characteristic T>C spectrum at ApT dinucleotides, was first discovered in liver cancers, and exhibits a strong transcriptional strand bias indicative for all damage occurring on adenines (Alexandrov et al., 2013b). Moreover, SBS16 is associated with alcohol and tobacco consumption, and appears in liver cancers predominantly in highly expressed genes, including *CTNNB1* (Letouzé et al., 2017).

Mutational signatures due to chemical compounds (SBS22/24)

SBS22 is dominated by T>A mutations and is most likely due to the exposure to aristocholic acid (Alexandrov et al., 2020), which is often the ingredient of traditional herbal remedies for various health problems ranging from arthritis to menstrual symptoms. Like many other

exogenous agents, the compound first needs to be metabolised, but can then form covalent adducts with adenosines (Poon et al., 2014).

Aflatoxins are byproducts of mould growing on food, which are metabolised to epoxide compounds that covalently bind to guanines, leading to G>T mutations (Poon et al., 2014). It is thought that the C>A rich spectrum of SBS24 represents exposure to aflatoxins (Alexandrov et al., 2020).

Defects in the base excision repair pathway introduce distinct C>T and C>A patterns (SBS30 and SBS36)

SBS30 is dominated by C>T mutations and was first characterised in a single sample of a breast cancer cohort (Nik-Zainal et al., 2016). Drost et al. (2017) were able to recover SBS30 in CRISPR-Cas9 knocked out *NHTL1* organoids, which are deficient in BER, as the *NTHL1* gene encodes a glycosylase (Sec. 1.4.2). Importantly, the authors were able to verify a germline nonsense mutation in the *NTHL1* gene in the sample where SBS30 was first discovered.

SBS36 represents another mutational signature that is linked to BER deficiency. However, SBS36 exhibits a spectrum dominated by C>A mutations and is linked to the inactivation of the *MUTYH* gene. The MUTYH protein plays an important role in the removal of 8-oxo-dG, as it scans newly synthesised DNA after replication and excises falsely incorporated adenines at A:8-oxo-dG mispairs (Viel et al., 2017).

Signatures of somatic hypermutation (SBS9, 84, 85)

Several signatures have been reported to be associated with the activity of AID, which is involved in the development of germinal B-cells. The enzyme deaminates cytosine to uracil during somatic hypermutation such that dependent on the chosen DNA repair pathway, different mutational patterns arise. Activation of the MMR pathway with recruitment of Pol η gives rise to the non-canonical AID pattern characterised by A>C or A>G mutations at WA motifs, and is represented by SBS9 (Alexandrov et al., 2013b). However, direct replication over an U:G lesion, or removal of U via the BER pathway followed by replication, induce the canonical AID signature characterised by a C>T or C>G mutations at WRCY motifs. A signature matching this nucleotide motif was found by Kasar et al. (2015), who extracted mutational signatures from clustered (and unclustered) single base substitutions. The canonical AID signature is represented by SBS84 and SBS85 in the COSMIC catalogue (Alexandrov et al., 2020).

1.7 Aims of this thesis

The discussion of mutagenesis in context of epigenetic regulation and fundamental cellular processes such as transcription and replication revealed a strong interplay between each of the individual subprocesses. In fact, none of them occur in isolation, and to make things even more complicated, they are often coupled to specific DNA repair mechanisms, such as transcription to NER, and replication to MMR. Additionally, there exist a multitude of different endogenous and exogenous mutagenic sources, each producing distinct mutational patterns in different cell or tissue types, thus introducing another layer of complexity.

In this thesis, I provide solutions to outstanding tasks in the field of mutational signature analysis. In particular, I aim to provide a novel framework to extract mutational signatures that

1. takes into account genomic heterogeneity due to cellular processes such as transcription and replication as well as the epigenome,

2. factors in the complete spectrum mutation types including MNV, indels, and structural variants,

3. and introduces methodological improvements by performing the inference with a robust noise model.

Chapter 2 shifts the focus to methodical aspects of mutational signature analysis. First, I will introduce the reader to non-negative matrix factorisation, the method that is at the heart of mutational signature analysis, and point out statistical considerations arising when NMF is applied to mutation count data. Sec. 2.2 introduces the methodology - TensorSignatures - I developed in the course of my PhD, assesses its properties, and compares the performance of the algorithm to previous approaches. The last section of this chapter explores novel approaches to deploy computationally intensive algorithms like TensorSignatures to cloud computing platforms, thus making the usage of scientific research pipelines more efficient and available to a broader community of users.

Chapter 3 showcases applications of TensorSignatures. In Sec. 3.1, I applied TensorSignatures to the PCAWG dataset which revealed the spectra and genomic properties of 20 tensor signatures. This analysis revealed two distinct mutational signatures of UV exposure found in active and quiescent chromatin, and revealed transcription-associated mutagenesis manifesting as A[T>C] mutations in a range of cancer types. TensorSignatures capability to incorporate other mutation types enable the algorithm to distinguish two APOBEC signatures reflecting highly clustered, double strand break repair initiated, and lowly clustered

replication-driven mutagenesis. Lastly, I demonstrate two signatures indicative for somatic hypermutation, producing a strongly clustered, TSS-associated signature in lymphoid cancers, which is distinct from a weakly clustered TLS signature found in multiple tumour types. Sec. 3.2 verifies these finding in a completely independently sequenced cohort of the Hartwig Medical Foundation (HMF), and Sec. 3.3 illustrates applications of TensorSignatures in a variety of datasets.

Chapter 4 concludes my thesis with a summary of findings, and a review of potential limitations of my research. Moreover, I will outline future perspectives and directions of mutational signature analysis that could further expand the understanding of the interplay between mutational processes and genomic features.

Chapter 2

Methods

The first chapter of this thesis introduced the genomic organisation of DNA, fundamental biological processes such as replication and transcription, and how these factors influence mutagenesis, as well as the great value of matrix-based mutational signature analysis as a tool to characterise mutational patterns in cancer genomes. However, there are limitations to this method, for example, it fails to account for the aforementioned genomic influences, does not support the concomitant inclusion of mutation types other than single nucleotide variants, and may be overly sensitive to statistical outliers. In this work, I addressed the shortcomings of current mutational signature analysis approaches by developing a novel non-negative *tensor* factorisation approach termed TensorSignatures. The key innovation of this method lies in extending the definition of a matrix (a two-dimensional array) to a tensor (a multi-dimensional array), which allows to categorise single nucleotide variants with respect to a multitude of different genomic annotations. Moreover, the method supports the incorporation of other mutation types, and employs a robust probabilistic model based on the negative binomial distribution.

These features enable TensorSignatures to characterise mutational processes in greater depth, for example, the method does not only yield signature spectra and corresponding exposures, but also quantifies the degree of strand bias with regard to transcription or replication, assesses the activity of signatures across distinct epigenetic regions, and reveals a signature's propensity to generate clustered mutations. Moreover, native inclusion of other mutation types provides a more thorough understanding by revealing the full mutational imprint of a mutational process, and imposing an appropriate statistical model enables to select the appropriate number of signatures through validated estimators rather than contentious heuristics. Finally, I implemented TensorSignatures using the modern machine learning library TensorFlow, which enables GPU-accelerated inference of large amounts

of data, thereby ensuring that the method is well-prepared for future applications when the number of available cancer genomes markedly exceeds the ten thousands mark.

This chapter will be divided into four main themes. First, I will discuss some conceptual limitations of non-negative matrix factorisation (NMF), the method underlying mutational signature analysis, and point out the challenges of modelling count data. Also, I will show how NMF models with general loss functions may be fitted using automatic differentiation and gradient descent algorithms on machine learning platforms such as Tensorflow and Pytorch (Sec. 2.1). Second, I will introduce in Sec. 2.2 the structure of the mutation count tensor, and dive deep into the parametrisation of the TensorSignature model. Third, I will assess the performance of the algorithm in simulation studies, and compare the inference to other approaches that have been utilised previously (Sec. 2.3). The last section 2.4 of this chapter explores possibilities to deploy research pipelines like TensorSignatures to cloud computing platforms, and illustrates this by introducing the TensorSignaturesOnline web application.

Contributions

This chapter is partly based on the bioarxiv manuscript "Learning Mutational Signatures and their Genomic Properties with TensorSignatures" by Harald Vöhringer, Arne van Hoeck, Edwin Cuppen and Moritz Gerstung. H.V. conducted all bioinformatic analyses and produced the figures. A.v.H. and E.C. curated HMF data and provided computing resources for HMF data analysis by H.V.. M.G. conceived and supervised the analysis and developed code for categorising mutations. H.V. and M.G. wrote the manuscript with input from A.v.H. and E.C.

2.1 Properties and limitations of non-negative matrix factorisation

Consider the NMF model (Eq. 1.1) used to decompose a catalogue of cancer genomes $\mathbf{C}$ into a set of mutational signatures $\mathbf{S}$ and their constituent activities or exposures $\mathbf{E}$, resulting in a natural and interpretable part-based representation of the data. While NMF refers to a general technique, most implementations minimise one of the following two objective functions over $\mathbf{S}$ and $\mathbf{E}$:

$$\text{Sum of squared errors objective:} \quad ||\mathbf{C} - \mathbf{SE}||^2 = \sum_i \sum_j (\mathbf{C}_{ij} - (\mathbf{SE})_{ij})^2 \tag{2.1}$$

or

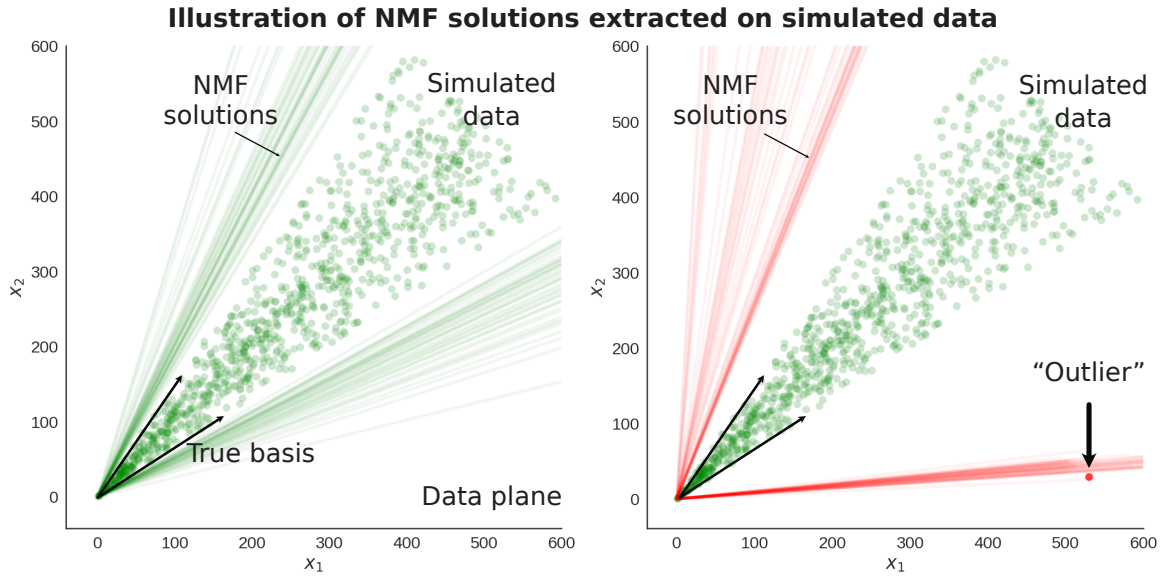


Fig. 2.1 **Illustration of NMF solutions extracted from a simulated dataset generated with 2-dimensional basis vectors.** Left panel: Illustration of the data plane including simulated data points, as well as the true and inferred basis vectors. Right panel: Addition of a single outlier introduces significant bias to extracted NMF solutions (here we used the standard sum of squared errors objective to compute depicted basis vectors).

$$\text{Divergence objective:} \quad D(\mathbf{C}||\mathbf{SE}) = -\sum_i \sum_j (\mathbf{C}_{ij} \log (\mathbf{SE})_{ij} - (\mathbf{SE})_{ij}). \tag{2.2}$$

To minimise these functions over $\mathbf{S}$ and $\mathbf{E}$, NMF deploys "multiplicative update rules" (Eq. 2.3), whose convergence may be proved using auxiliary functions (Lee and Seung, 2000).

$$\mathbf{E}_{kj} \leftarrow \mathbf{E}_{kj} \frac{(\mathbf{S}^T\mathbf{C})_{kj}}{(\mathbf{S}^T\mathbf{SE})_{kj}}, \quad \mathbf{S}_{ik} \leftarrow \mathbf{S}_{ik} \frac{(\mathbf{CE}^T)_{ik}}{(\mathbf{SEE}^T)_{ik}} \tag{2.3}$$

Particularly, the algorithm initialises $\mathbf{S}$ and $\mathbf{E}$ with random non-negative values, and minimises the objective function iteratively, by optimising the values of the first matrix $\mathbf{E}$, and using these updated values to calculate the next set of values in the second matrix $\mathbf{S}$. This process continues until the algorithm reaches a local minimum of the objective function.

2.1.1 The geometry of NMF solutions

NMF extracts a set of non-negative basis vectors forming a simplical cone, whose linear combinations span the cloud of data points of the input matrix. To illustrate this, I simulated 1,000 data points (green dots in Fig. 2.1) by computing random non-negative linear

combinations of two linearly independent vectors (true basis vectors, black arrows in Fig. 2.1): $\binom{6}{4}\binom{4}{6}$. Next, I applied NMF with different initial values to generate 50 solutions of rank two (green rays in Fig. 2.1). Note how these rays surround the cloud of data points implying that NMF solutions are based upon the set of points which make up the border of the data cloud. It is important to keep that in mind when applying the algorithm, because it indicates that outliers, i.e. data points deviating substantially from the data cloud, may affect NMF solutions considerably. To illustrate this, I used the same simulated data set, but added such an outlier (red point in Fig. 2.1). Applying NMF to this data yielded a substantially different set of basis vectors (red rays in Fig. 2.1). This shows that the presence of just one unusual data point may affect the solutions provided by NMF. For this reason, it seems to be mandatory to use appropriate noise models for different types of input data.

2.1.2 The probabilistic interpretation of NMF

From a probabilistic point of view, minimising the sum of squared errors objective implies to impose a Gaussian distribution to the count data in $\mathbf{C}$

$$\mathbf{C}_{ij} \sim \mathcal{N}((\mathbf{SE})_{ij}, \sigma^2). \tag{2.4}$$

However, base substitution counts are clearly non-continuous and always ≥ 0. Applying NMF to count data as outlined in Eq. 2.3, may yield improbable solutions since it applies incorrect model assumptions. In context of count data, the divergence objective is more appropriate, since the maximum likelihood interpretation of $D(\mathbf{C}||\mathbf{SE})$ implies

$$\mathbf{C}_{ij} \sim \text{Pois}((\mathbf{SE})_{ij}), \quad \text{where Pois}(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda},\ x \in \{0,1,2,\dots\}. \tag{2.5}$$

Although the Poisson distribution represents a reasonable choice for modelling single base substitution counts, it implies certain limitations. Most importantly, the Poisson model assumes *equidispersion*, i.e. the variance to mean ratio equals one:

$$\text{if } X \sim \text{Pois}(\lambda), \quad \text{then} \quad \mathbb{E}[X] = \lambda, \text{and } \mathbb{V}[X] = \lambda. \tag{2.6}$$

Therefore, when counts are larger on average, they also tend to be more variable. Close inspection of the Poisson density reveals that as the mean λ increases, the skew decreases and the distribution becomes more bell-shaped. Note that this is crucially different to the Normal model where the variance σ^2 is assumed to be constant (Eq. 2.4).

2.1.3 Modelling overdispersion

Count data often varies more than we would expect if the generative process was truly Poisson. If variances are larger than corresponding means, count data is classified as *overdispersed*. A common cause for overdispersion is heterogeneity among subjects, which is very likely to be observed in SNV count data, because cancer patients may be exposed to a multitude of different exogenous and endogenous mutagenic factors. For example, cancers with mutations in the exonuclease domain of Pol ε often exhibit a hypermutator phenotype, characterised by exceedingly large numbers of mutations in comparison to other cancers.

There are probabilistic models that can account for overdispersion, one of them is the so called negative binomial (NB) distribution. A useful way to think about this distribution is that it represents a generalisation of the Poisson distribution where the mean parameter λ (see Eq. 2.5) is itself a Gamma distributed random variable

$$\mathrm{NB}(x|\mu,\tau) = \frac{\Gamma(\tau+x)}{\Gamma(\tau)x!}\left(\frac{\tau}{\tau+x}\right)^{\tau}\left(\frac{\mu}{\tau+\mu}\right)^{x}. \tag{2.7}$$

Equation 2.7 tells us that the NB distribution has an additional non-negative dispersion parameter τ, which controls the variance of the distribution. Also, observe that

$$\text{if } X \sim \mathrm{NB}, \quad \text{then} \quad \mathbb{E}[X] = \mu, \text{ and } \mathbb{V}[X] = \mu + \frac{1}{\tau}\mu^2, \tag{2.8}$$

implying that the NB distribution converges to a Poisson model as $\tau \to \infty$, while small values of τ enable to scale the degree of overdispersion relative to the Poisson.

Signs for overdispersion in SNV count data

In an attempt to estimate the heterogeneity of single base substitution counts, we separated the SNV spectrum of skin melanoma samples by chromosome to obtain internal replicates. We denote the count for single base substitution type i on chromosome j of sample k as C_{ijk}. Under the simplifying assumption that mutational processes do not have preferences among chromosomes, we would expect that the mean single base substitution count i in sample k, denoted as $\hat{C}_{ik}$, is roughly proportional to the trinucleotide frequency of chromosome j

$$\hat{C}_{ik} = \sum_{j=1}^{22} C_{ijk} \times F_{ij} \tag{2.9}$$

where F_{ij} is the frequency of trinucleotide i on chromosome j. To estimate the degree of dispersion, we used the estimator $\hat{C}_{ik}$ to fit C_{ijk} with a generalised linear model using negative binomial regression. In our analysis, we found $\tau \approx 50$ indicating that base substitution count

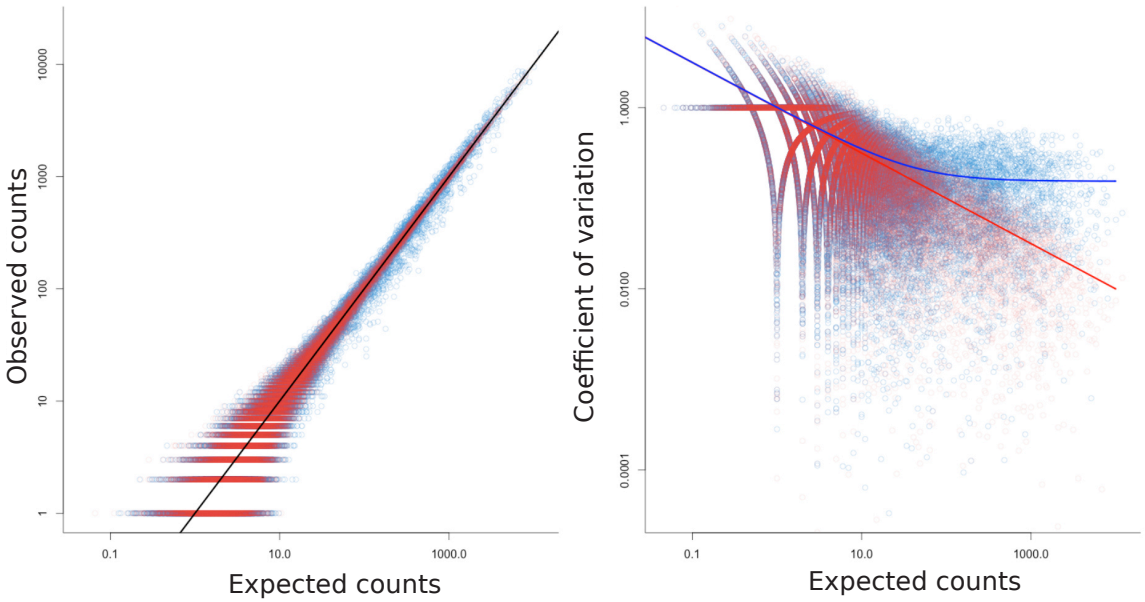


Fig. 2.2 **Evidence for overdispersion in base substitution count data.** The left panel shows how expected base substitution counts correlate with observations (blue points), and simulated observations drawn from a Poisson distribution (red points). The right plot illustrates how the standard deviation relates to the mean ($CV = \frac{\sigma}{\bar{x}}$) where blue points show actual data, and red points simulated data sampled from a Poisson distribution. To estimate the degree of overdispersion, we fitted a generalised linear model using NB regression (blue line) and Poisson regression (red line). In this example, we used skin melanoma samples from the PCAWG dataset, and used single base substitution counts from diploid chromosomes.

data from skin melanoma samples are highly overdispersed (Fig. 2.2). This situation makes it difficult for NMF to extract representative signatures, because the method naturally selects "corner points" of the data to define the basis of its low rank representation (Sec. 2.1.1). There is a danger to find mutational signatures based on particular samples, rather than spectra that are truly representative for a mutagenic biological process.

Examples from the literature indicating overfitting of mutational signatures

We found examples in the literature that are symptomatic for this issue. For example, Hayward et al. (2017) report the identification three novel UV signatures. Although the authors justify the existence of these signatures by proposing reasonable underlying mutational processes, it remains to be proved whether these spectra are truly due to differential types of UV damage (in this case cyclobutane pyrimidine dimers and indirect DNA damage after ultraviolet radiation). Also, we found that there is at least one sample in their dataset in which the contribution of these UV-signatures is almost at 100 %, which indicates that these samples

may represent signature-defining outliers, fooling the algorithm to fit distinct signatures for particular samples, rather than to identify truly prototypical mutational patterns.

2.1.4 Fitting NMF models with automatic differentation and gradient descent

The simple formulation of the sum of squared errors and the divergence objective in NMF enable the definition of simple iterative update rules based on coordinate descent (e.g. Eq. 2.3). However, the probabilistic implications of these functions render such NMF models to be inappropriate for certain types of data. To nevertheless perform NMF with more realistic error models, modern machine learning libraries such as Tensorflow provide a suitable framework for easy implementation and execution.

TensorFlow: A library for defining computational graphs

Tensorflow is a versatile open-source end-to-end platform with the goal to facilitate the creation and deployment of machine learning models. Internally it is a symbolic math library that defines dataflows and uses differential programming, i.e. Tensorflow allows to define a series of differentiable operations in specific order on data, making the framework particularly suitable for applications such as neural networks. The library was initially developed by GoogleBrain for internal Google use, but released to the public in 2015 under the Apache License 2.0 (Abadi et al., 2016).

Tensorflow can be considered as a framework for defining computational graphs that describe certain computations. A graph is a data structure consisting of nodes and edges, which in Tensorflow describe local units of computation, and input and outputs, respectively. In other words, in a Tensorflow computation graph, nodes represent *operations*, and the values flowing along edges are *tensors*. An operation takes tensors as input and produces tensors as output, and may represent common mathematical operation such as addition or matrix multiplication, but also more complicated transformations such as convolutions. On the other hand, tensors naturally represent the inputs or outputs of operations, for example, a matrix multiplication takes two 2D tensors as input, and outputs another 2D tensor (Abadi et al., 2016).

Two other important concepts in Tensorflow are *variables* and *automatic differentiation*. A TensorFlow variable is a special type of operation that contains an internal state, i.e. a tensor, whose value(s) can be changed by running operations on it. Specific operations allow to read and modify the values of a variable, making them suitable to store model parameters in machine learning models. Automatic differentiation is a set of routines to numerically

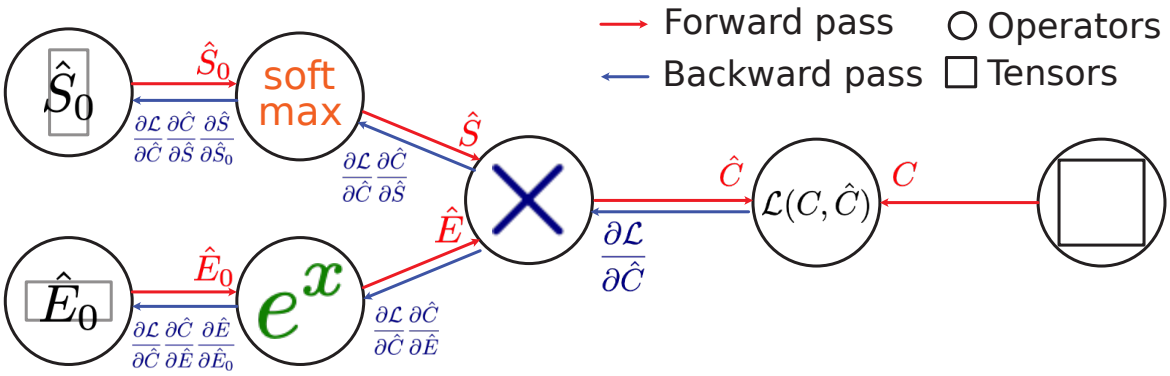


Fig. 2.3 **A computational graph to compute a non-negative matrix factorisation in TensorFlow.** A computational graph contains nodes and edges. In TensorFlow, these correspond to operations and tensors, respectively.

evaluate the derivative of a function, which enables each node in a Tensorflow graph, to compute its derivative with respect to its successor node (Abadi et al., 2016).

Training machine learning models in TensorFlow

To create a machine learning model in TensorFlow, users define a directed acyclic computational graph of model parameters and layers that terminate with a loss function. A layer maybe considered as a composite of mathematical operations acting on the variables of the model, while a loss function is a scalar function, quantifying the difference between the predicted value of the model and the ground truth. To train a learning model, Tensorflow will update the model parameters iteratively by computing forward and backward passes. The forward pass subsumes all steps required to compute the prediction as defined in the computational graph, and terminates with the layer that computes the objective function. In contrast, the backward pass updates model parameters outgoing from the layer computing the loss, by moving them in the direction that maximally decreases the output of the objective function. To achieve this, Tensorflow uses backpropagation, an algorithm which makes use of automatic differentiation to compute the gradient of the loss with respect to each parameter using the chain rule, i.e. it computes the gradient one layer at a time by iterating from the last layer to the first. To update the parameters, Tensorflow makes use of gradient descent, which simply subtracts a scaled[1] version of this gradient from the current parameter value (Abadi et al., 2016).

[1]Scaling gradients is important to avoid numerical problems due to overflowing gradients, and determines how fast the model learns. This is usually controlled via a hyperparameter termed “learning rate”.

A computational graph to compute a NMF using the Tensorflow framework

Given the propensity of NMF to be highly affected by outliers (Sec. 2.1.1), and the likeliness to find overdispersion in SNV count data (Sec. 2.1.3), I propose to compute the NMF based on the NB distribution. If we assume that a single base substitution count $\mathbf{C}_{ij}$ is a NB distributed random variable generated by a fixed number of mutational signatures $\mathbf{S}$ and their corresponding exposures $\mathbf{E}$, we may formulate the matrix decomposition as

$$\mathbf{C}_{ij} \sim \mathrm{NB}((\mathbf{SE})_{ij}, \tau) \quad \text{where } \mathbf{C} \in \mathbb{N}_0^{p\times n}, \mathbf{S} \in \mathbb{R}_+^{p\times s}, \text{and } \mathbf{E} \in \mathbb{R}_+^{s\times n}. \tag{2.10}$$

To implement a NMF using a computational graph, one may consider to apply non-linear functions such as the softmax and exponential function to the signature and exposure matrix $\mathbf{S}$ and $\mathbf{E}$, thus ensuring they remain positive and satisfy constraints for mathematical identifiability. After matrix multiplying $\mathbf{S}$ and $\mathbf{E}$ to obtain a prediction for the true $\mathbf{C}$, one simply minimises the negative NB log-likelihood

$$\log \mathcal{L}(\mathbf{C}; \mathbf{S}, \mathbf{E}, \tau) = -\sum_i \sum_j \mathbf{C}_{ij} \log (\mathbf{SE})_{ij} - (\tau + \mathbf{C}_{ij}) \log (\tau + (\mathbf{SE})_{ij}). \tag{2.11}$$

by iteratively computing forward and backward passes until the weights of $\mathbf{S}$ and $\mathbf{E}$ reach a local minimum. To illustrate how simple it is to achieve this using Tensorflow, I attached the code that implements the graph from Fig. 2.3.

```
import tensorflow as tf
p, s, n = 96, 10, 100
C = tf.constant(counts)
E0 = tf.Variable( tf.random.normal(p, s) )
S0 = tf.Variable( tf.random.normal(s, n) )
E = tf.exp(E0)
S = tf.softmax(S0, dim=0)
Chat = tf.matmul(S, E)
loss = log_likelihood(C, Chat)
```

Together with the benefit of having native GPU support, which enables a massive speed up of tensor computations, we decided to implement the TensorSignature software using the Tensorflow library.

2.2 TensorSignatures

I developed TensorSignatures with the goal of creating a software to

1. characterise mutational processes with respect to various genomic properties at the stage of signature definition,
2. define mutational processes in terms of their complete mutational imprint, and
3. introduce algorithmic improvements making the software less prone to fit spurious mutational signatures.

In this section, I will explore the organisation of TensorSignatures' input data (Sec. 2.2.1), and explain the parameterisation of the model, allowing it to learn a mutational signature's properties and other mutation types (Sec. 2.2.2-2.2.6).

2.2.1 Multi-dimensional input data

The prime objective of TensorSignatures suggests to partition mutation counts with respect to different genomic features. Such data may be represented in form of a multidimensional tensor, or alternatively, as a flattened matrix. While the former implies to perform a tensor factorisation, for which currently no out-of-the-box solution exists, the latter hardly scales with increasing numbers of genomic states. Consider two genomic dimensions, for example, transcription and replication, each with three states, i.e. coding, template and unknown transcription strand, and leading, lagging and unknown replication strand. Representing such data in a flattened array would require a $(3 \cdot 3 \cdot 96) \times n$ matrix, thus requiring to fit 864 (!) parameters per signature (Eq. 1.1). In contrast, a tensor representation allows to parametrise the model such that the number of required parameters per signatures scale additively with the number of genomic states, thus making a characterisation of mutational signatures with respect to multiple genomic states possible.

TensorSignatures receives two inputs: $\mathbf{C}^{\mathbf{SNV}}$ and $\mathbf{C}^{\mathbf{other}}$

To enable the genomic characterisation and the concurrent discovery of mutational signatures with respect to their full mutational imprint, TensorSignatures receives two data inputs: a mutation count tensor $\mathbf{C}^{\mathrm{SNV}}$, whose multidimensional structure enables to annotate SNVs with various genomic features, including

- transcription and replication strand,

- epigenetic state,
- nucleosomal position,
- and clustering state,

and a mutation matrix containing the counts of all other mutation types $\mathbf{C}^{\text{other}}$.

The SNV mutation count tensor

Similar to conventional signature analysis (Alexandrov et al., 2013b), we classify SNVs with respect to their trinucleotide context, but count pyrimidine ([C/T>·]) and purine ([A/G>·]) mutations separately, resulting in twice the number of features (192 rather than 96)[2]. However, rather than appending purine mutation types to the SNV count matrix (i.e $\mathbf{C}^{\text{SNV}} \in \mathbb{N}_0^{192\times n}$ where n denotes the number of samples), we include these additional features by introducing a new dimension to the array $\mathbf{C}^{\text{SNV}} \in \mathbb{N}_0^{2\times p\times n}$, where $p = 96$. This structure may be understood as two stacked matrices $\mathbf{C}^{\text{SNV}}_{\text{pyr}} \in \mathbb{N}_0^{p\times n}$ and $\mathbf{C}^{\text{SNV}}_{\text{pur}} \in \mathbb{N}_0^{p\times n}$ as shown in Eq. 2.12. Note the organisation of the tensor, which matches the indices of pyrimidine mutation types to their respective purine reverse complements.

$$
\begin{array}{r}
\\
1\ \text{A[C}\!\rightarrow\!\text{A]A}\ \{ \\
2\ \text{A[C}\!\rightarrow\!\text{A]C}\ \{ \\
3\ \text{A[C}\!\rightarrow\!\text{A]G}\ \{ \\
\\
96\ \text{T[T}\!\rightarrow\!\text{G]T}\ \{
\end{array}
\overbrace{
\begin{array}{r}
1\ \text{T[G}\!\rightarrow\!\text{T]T}\ \{ \\
2\ \text{G[G}\!\rightarrow\!\text{T]T}\ \{ \\
3\ \text{C[G}\!\rightarrow\!\text{T]T}\ \{ \\
\\
\\
\end{array}
\left[\begin{array}{cccc}
C_{111} & C_{112} & \dots & C_{11n} \\
C_{121} & C_{122} & \dots & C_{12n} \\
C_{131} & C_{132} & \dots & C_{13n} \\
\vdots & \vdots & \ddots & \vdots \\
C_{1p1} & C_{1p2} & \dots & C_{1pn}
\end{array}\right]}^{\mathbf{C}^{\text{SNV}}_{\text{pyr}}}
\overbrace{\left[\begin{array}{cccc}
C_{211} & C_{212} & \dots & C_{21n} \\
C_{221} & C_{222} & \dots & C_{22n} \\
C_{231} & C_{232} & \dots & C_{23n} \\
\vdots & \ddots & & \vdots \\
C_{2p2} & \dots & & C_{2pn}
\end{array}\right]}^{\mathbf{C}^{\text{SNV}}_{\text{pur}}}
\tag{2.12}
$$

Transcription directionality To assign single base substitutions to template and coding strand, we partitioned the genome by transcription directionality (trx(+)/trx(-)) using GENCODE v19 definitions. The transcription machinery synthesises RNA always in 5' to 3' direction, implying that template and coding strand of trx(-) genes are 5' to 3' and 3' to 5'

[2]Classifying pyrimidine and purine base substitution types is at this stage pure convenience, as it allows to assign transcription and replication directionality directly to each mutation. The redundancy in the representation of mutation types is later resolved by adding purine mutation types to their pyrimidine equivalents from coding and template strands, and leading and lagging strands.

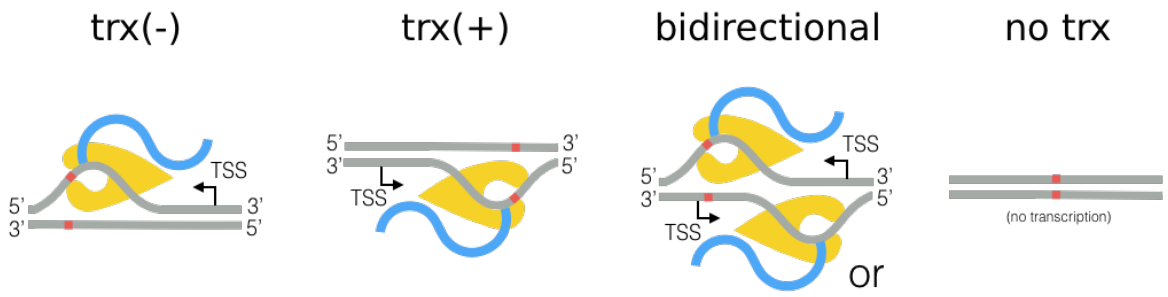


Fig. 2.4 **Annotating SNVs with transcription directionality.** To annotate SNVs with template and coding strand annotation, we partitioned the genome into non-overlapping trx(+) and trx(-) regions, which indicate transcription from the non-reference or reference strand, respectively. If a genomic region did not contain a gene, or if genes are present on both reference and non-reference strand, we assigned a * (star) to this region (removing bidirectionally transcribed genomic regions reduces the percentage of the transcribed genome from 53 % to 47 %).

oriented, and vice versa for trx(+) genes (Fig. 2.4). Since SNVs are called on the 5' to 3' strand of the DNA (i.e. + strand of the reference genome), we can unambiguously determine whether the pyrimidine of the mutated Watson-Crick base pair was on the coding or template strand. For example, a G>A substitution in a trx(-) gene corresponds to a coding strand C>T mutation, because transcription directionality dictates that the mutated G sits on the template strand. Splitting SNVs in this fashion requires to introduce an additional dimension of size three (coding, template and unknown strand) to the count tensor $\mathbf{C}^{\text{SNV}} \in \mathbb{N}_0^{3\times 2\times p\times n}$ (where $p = 96$ and n is the number of samples).

Replication directionality To assign single base substitutions to leading and lagging strand, we leveraged Repli-seq data from the ENCODE consortium (Hansen et al., 2010; Thurman et al., 2012), which map the sequences of nascent DNA replication strands throughout the whole genome during each cell cycle phase. Repli-seq profiles relate genomic coordinates to replication timing (early and late), where local maxima (peaks) and minima (valleys) correspond to replication initiation and termination zones. Regions between those peaks and valleys are characterised by steep slopes, whose algebraic sign (rep(+) or rep(-)) indicate whether the leading strand is replicated to the left or right direction (left or right replication) when the DNA is viewed in standard orientation (Fig. 2.5).

To partition the genome into non-overlapping right and left replicating regions, we computed the mean of slopes from Repli-seq profiles of five cell lines (GM12818, K564, Hela, Huvec and Hepg2) using finite differences (Koren et al., 2012). We marked regions with a plus (+) if the slope was positive (and therefore left-replicating), and with minus

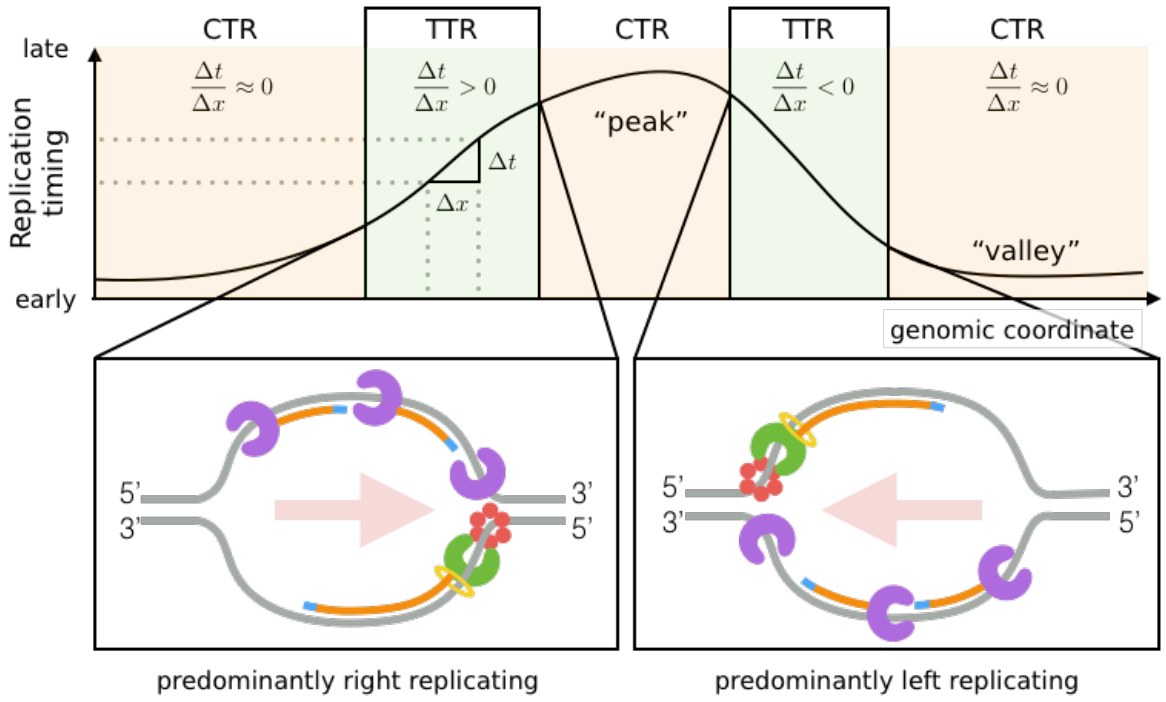


Fig. 2.5 **Partitioning the genome by replication direction.** The upper panel depicts a schematic RepliSeq profile with peaks and valleys known as constant timing regions (CTRs), which comprise predominantly early and late firing ORIs, respectively. On the other hand, regions between peaks and valleys, transition timing regions (TTRs), are characterised by a slope of large magnitude in which the algebraic sign (+/-) indicates whether the region is mostly replicated to the right (+) or left direction (-) when DNA is viewed in standard orientation. In genomic regions where the slope of the RepliSeq curve has a large magnitude, the directionality of the replication bubble can be determined (green leading strand polymerase, purple lagging strand polymerase).

(-) if the slope was negative (and hence right-replicating). To confidently assign these states, we required that the absolute value of the mean of slopes was at least larger than two times its standard deviations, otherwise we assigned the unknown (*) state to the respective region. Consensus annotations enable to annotate 24.9 % of the genome (in comparison to approximately 38% of the genome which may be achieved with tissue specific repliseq annotations). Using this convention, for example, a C>A variant in a rep(+) region corresponds to a template C for leading strand DNA synthesis (and a template G for lagging strand). Subsequent assignment of single base substitutions to leading and lagging strand is analogous to the procedure we used for transcription strand assignment, and adds another dimension of size of three to the count tensor ($\mathbf{C}^{\text{SNV}} \in \mathbb{N}_0^{3\times3\times2\times p\times n}$).

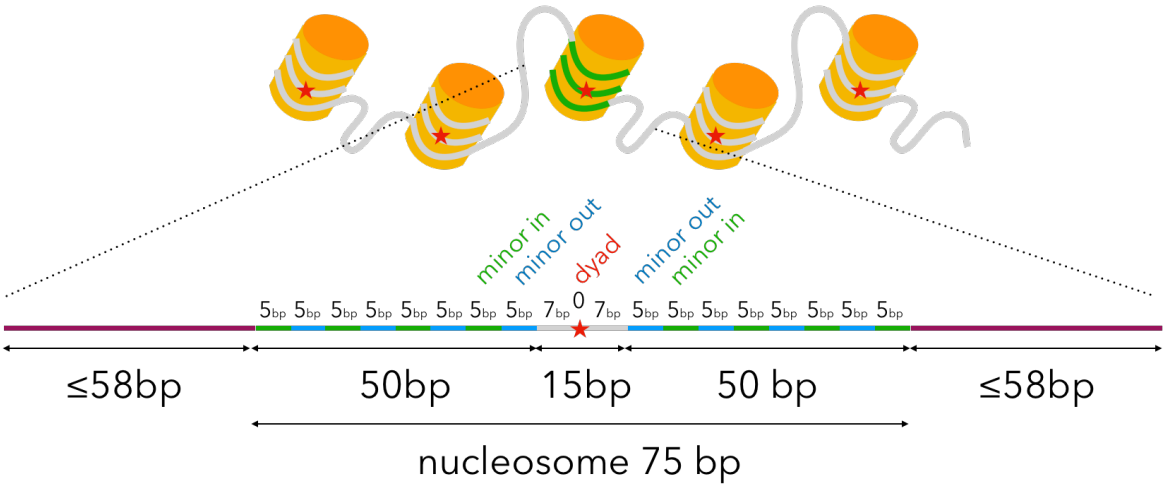


Fig. 2.6 **Nucleosomal states mark minor grooves facing towards, and away from histones, and linker regions.** We used the coordinates of DNA dyads from MNase cut efficiency experiments to annotate nucleosomal DNA with annotations, marking the minor grooves facing away and towards histone proteins, and linker DNA (purple).

Nucleosomal states Although nucleosomal DNA binding is mostly non-sequence specific, histone bound DNA often features AT-rich followed by GC-rich minor grooves, which have the propensity to bend the molecule favourably, thus facilitating histone interactions. For this reason, nucleosomal DNA organisation may lead to differential susceptibility to mutational processes (see also Sec. 1.1, Fig. 1.1). To assign single base substitutions to minor grooves facing away from and towards histones, and linker regions in between nucleosomes, we used nucleosome dyad (midpoint) positions of human lymphoblastoid cell lines mapped in MNase cut efficiency experiments (Pich et al., 2018). We partitioned nucleosomal DNA by first adding 7 bp to both sides of a dyad, and assigning to the following 50 bp alternating 5 bp minor out and minor in DNA stretches, followed by a linker region with a maximum of 58 bp (Fig. 2.6). Subsequent assignment of SNVs to these states adds another dimension of size four to the count tensor ($\mathbf{C}^{\text{SNV}} \in \mathbb{N}_0^{3\times3\times4\times2\times p\times n}$).

Epigenetic states To assign single base substitutions to different epigenetic environments, we used functional annotations from the 15-state ChromHMM model provided by the Roadmap epigenomics consortium (Ernst and Kellis, 2012; Kundaje et al., 2015), which integrates multiple chromatin datasets such as ChIP-seq data of various histone modifications. To find state annotations that are robust across all cancer tissues, we defined an epigenetic consensus state by combining state annotations from 127 different Roadmap cell lines. Here, we required that at least 70 % of the cell lines agreed in the Chrom-HMM state to accept the

state for a given genomic region. Partitioning SNVs by Chrom-HMM states adds another dimension of size 16 to the count tensor ($\mathbf{C}^{\text{SNV}} \in \mathbb{N}_0^{3\times3\times16\times4\times2\times p\times n}$).

Clustering states To identify clustered single base substitutions, we used inter mutation distances (Y_k in bp) between consecutive mutations on a chromosome as observations for a two state (X_k = {clustered, unclustered}) Hidden Markov Model (HMM) with initial/transition distribution

$$p_{X_1}(x_1) = \begin{cases} 0.01 & \text{if } x_1 \text{ clustered} \\ 0.99 & \text{if } x_1 \text{ unclustered} \end{cases} \quad p_{X_{k+1}|X_k}(x_{k+1}|x_k) \begin{cases} 0.99 & \text{if } x_{k+1} = x_k \\ 0.01 & \text{if } x_{k+1} \neq x_k \end{cases} \tag{2.13}$$

and observation distribution

$$p_{Y_k|X_k}(y_k|x_k) = \begin{cases} \text{Geom}(p = 1/100) & \text{if } x_k \text{ clustered} \\ \text{Geom}(p = (\frac{1}{n}\sum_{k=1}^{n} y_k)^{-1}) & \text{if } x_k \text{ unclustered} \end{cases} \tag{2.14}$$

We then computed the maximum a posteriori (MAP) state using the Viterbi algorithm to assign to each mutation the state clustered or unclustered, respectively. Adding clustered SNVs to the count tensor introduces another dimension to the count tensor ($\mathbf{C}^{\text{SNV}} \in \mathbb{N}_0^{3\times3\times16\times4\times2\times2\times p\times n}$).

Collapsing the pyrimidine/purine dimension So far, the structure of the SNV mutation count tensor embodies redundancy with regard to the assignment of transcription and replication directionality to pyrimidine and purine mutation types. For example, the count tensor explicitly contains the number of G>A and C>T substitutions on trx(-) genes, which correspond to C>T mutations on a coding and template strand respectively if both variant types are expressed solely in terms of pyrimidines. However, this redundancy is easily resolved by adding the purine mutation types from corresponding trx(+)/trx(-), and rep(+)/rep(-) to their equivalent pyrimidine mutations on template/coding and leading/lagging strand. This operation reduces the dimensionality of the final SNV count tensor $\mathbf{C}^{\text{SNV}} \in \mathbb{N}_0^{3\times3\times16\times4\times2\times p\times n}$.

Other mutation types

The second input for TensorSignatures is a 2-dimensional matrix $\mathbf{C}^{\text{other}} \in \mathbb{N}_0^{q\times n}$ containing the counts of q other mutation types, i.e. multinucleotide variants, deletions and insertions, and structural variants. While the data organisation of single base substitutions has to follow the tensor conventions outlined in the previous section, requirements for $\mathbf{C}^{\text{other}}$ are relaxed

and hence any mutation count matrix is accepted. The TensorSignatures' package implements routines to classify multinucleotide variants, deletions and insertions by type and length (see Tab. B.2, B.3), but does not provide an automated programme to catalogue structural variants. Representing other mutation types in a matrix (rather than a tensor) is motivated by the fact that it is difficult to partition SVs, MNVs and larger indels (> 1bp) variants uniquely to genomic features, as they may overlap with multiple genomic states.

2.2.2 The signature tensor

Recall that conventional mutational signature analysis uses NMF to decompose a catalogue of cancer genomes $\mathbf{C}$ to a set of mutational signatures $\mathbf{S}$ and their constituent activities or exposures $\mathbf{E}$ (Eq. 1.1). Similarly, TensorSignatures identifies a low dimensional representation of a mutation count tensor, but decomposes it to mutational spectra for coding and template strand, leading and lagging strand, and signature-specific multiplicative factors quantifying the propensities of mutational processes within specific genomic contexts. To enable strand specific extraction of mutational spectra requires to increase the dimensionality of the $p \times s$ sized signature matrix. To understand this, consider that two $p \times s$ matrices are at least needed to represent spectra for coding (C) and template (T) strand, suggesting a three dimensional $(2 \times p \times s)$ signature representation. Our model, however, also considers replication, which adds another dimension of size two for leading (L) and lagging (G) strand, and thus we represent mutational spectra in the four dimensional core signature tensor $\mathbf{T_0} \in \mathbb{R}^{2\times 2\times p\times s}$

$$\mathbf{T_0} = \begin{bmatrix} \mathbf{T_0}^{\mathrm{C/L}} & \mathbf{T_0}^{\mathrm{C/G}} \\ \mathbf{T_0}^{\mathrm{T/L}} & \mathbf{T_0}^{\mathrm{T/G}} \end{bmatrix} \quad \text{where } \mathbf{T_0}^{\mathrm{C/L}}, \mathbf{T_0}^{\mathrm{C/G}}, \mathbf{T_0}^{\mathrm{T/L}}, \mathbf{T_0}^{\mathrm{T/G}} \in \mathbb{R}_+^{p\times s}. \tag{2.15}$$

The mutation spectra $\mathbf{T_0}^{\cdot/\cdot}$ are normalised to 1 for each signature s, i.e., $\sum_{i=1}^{p}(\mathbf{T_0}^{\cdot/\cdot})_{is} = 1 \; \forall \; s$. However, the mutation count tensor also contains mutations from genomic regions for which strand assignment was not applicable. To still use these data for the factorization, we map such counts to a linear combinations of $\mathbf{T_0}$'s sub matrices. This is enabled by *stacking* strand specific $p \times s$ matrices of the core signature tensor, thereby forming linear combinations. For example, coding strand mutations for which replicational strand assignment was not applicable, are mapped to a linear combination of both coding strand specific sub matrices $\mathbf{T_0}^{\mathrm{C/L}}$ and $\mathbf{T_0}^{\mathrm{C/G}}$. Stacking sub matrices of $\mathbf{T_0}$ results in $\mathbf{T_1} \in \mathbb{R}_+^{3\times 3\times p\times s}$

$$\mathbf{T_1} = \begin{bmatrix} \mathbf{T_0}^{\mathrm{C/L}} & \mathbf{T_0}^{\mathrm{C/G}} & \frac{1}{2}(\mathbf{T_0}^{\mathrm{C/L}} + \mathbf{T_0}^{\mathrm{C/G}}) \\ \mathbf{T_0}^{\mathrm{T/L}} & \mathbf{T_0}^{\mathrm{T/G}} & \frac{1}{2}(\mathbf{T_0}^{\mathrm{T/L}} + \mathbf{T_0}^{\mathrm{T/G}}) \\ \frac{1}{2}(\mathbf{T_0}^{\mathrm{C/L}} + \mathbf{T_0}^{\mathrm{T/L}}) & \frac{1}{2}(\mathbf{T_0}^{\mathrm{C/G}} + \mathbf{T_0}^{\mathrm{T/G}}) & \mathbf{T_0}^{\mathrm{avg.}} \end{bmatrix} \tag{2.16}$$

where $\mathbf{T}_{\mathbf{0}}^{\text{avg}} = \frac{1}{4}(\mathbf{T}_{\mathbf{0}}^{\text{C/L}} + \mathbf{T}_{\mathbf{0}}^{\text{T/L}} + \mathbf{T}_{\mathbf{0}}^{\text{C/G}} + \mathbf{T}_{\mathbf{0}}^{\text{T/G}})$.

Tensor factors

We use the term *tensor factor* for variables of the model that are factored into the signature tensor to quantify different genomic properties of a mutational signature. The key idea is to express a mutational process in terms of a product of strand-specific spectra and a set of scalars, which modulate the magnitude of spectra dependent on the genomic state combination presented in the count tensor. However, to understand how tensor factors enter the factorisation, it is necessary to introduce the concept of broadcasting, which is the process of making tensors with different shapes compatible for arithmetic operations.

Broadcasting It is important to realise that it is possible to increase the number of dimensions of a tensor by prepending their shapes with ones. For example, a three dimensional tensor $\mathbf{X}$ of shape $\mathbb{R}_{+}^{2\times3\times5}$ has 2 rows, 3 columns and a depth of 5. However, we could reshape $\mathbf{X}$ to $\mathbb{R}_{+}^{1\times2\times1\times3\times5}$, or $\mathbb{R}_{+}^{2\times3\times5\times1}$, which would eventually change the order of values in the array, but not its content. These extra (empty) dimensions are called singletons or degenerates, and are required to make entities of different dimensionality compatible for arithmetic operations via *broadcasting*. Consider the following example

$$\underset{\mathbb{R}^{1\times2}}{\begin{bmatrix}1 & 2\end{bmatrix}} \odot \underset{\mathbb{R}^{2\times1}}{\begin{bmatrix}3\\4\end{bmatrix}} = \underbrace{\begin{bmatrix}1 & 2\\1 & 2\end{bmatrix} \cdot \begin{bmatrix}3 & 3\\4 & 4\end{bmatrix}}_{\text{broadcasting and element-wise multiplication } \cdot} = \underset{\mathbb{R}^{2\times2}}{\begin{bmatrix}3 & 6\\4 & 8\end{bmatrix}}. \tag{2.17}$$

The $\odot$ operator first copies the elements along their singleton axes such that the shape of both resulting arrays match, and then performs element-wise multiplication as indicated by the $\cdot$ symbol. This concept is similar to the tensor product $\otimes$ for vectors, but also applies to higher dimensional arrays, although this requires to define the shapes of all tensors carefully. For example, if $\mathbf{F} \in \mathbb{R}^{2\times2}$ and $\mathbf{H} \in \mathbb{R}^{1\times1\times3}$ then $\mathbf{F} \odot \mathbf{H}$ is an invalid operation, however, if $\mathbf{G} \in \mathbb{R}^{2\times2\times1}$, then $(\mathbf{G} \odot \mathbf{H}) \in \mathbb{R}^{2\times2\times3}$ is valid. Also, note that such operations are not necessarily commutative.

Transcriptional and replicational strand biases To quantify spectral asymmetries in context of transcription and replication, we introduce two vectors $\mathbf{b_t}, \mathbf{b_r} \in \mathbb{R}_{+}^{1\times s}$, stack and reshape them such that the resulting bias tensor $\mathbf{B} \in \mathbb{R}_{+}^{3\times3\times1\times s}$,

$$\mathbf{B} = \begin{bmatrix} \mathbf{b_r} \cdot \mathbf{b_t} & \mathbf{b_r^{-1}} \cdot \mathbf{b_t} & \mathbf{1} \cdot \mathbf{b_t} \\ \mathbf{b_r} \cdot \mathbf{b_t^{-1}} & \mathbf{b_r^{-1}} \cdot \mathbf{b_t^{-1}} & \mathbf{1} \cdot \mathbf{b_t^{-1}} \\ \mathbf{b_r} \cdot \mathbf{1} & \mathbf{b_r^{-1}} \cdot \mathbf{1} & \mathbf{1} \cdot \mathbf{1} \end{bmatrix}, \tag{2.18}$$

matches the shape of $\mathbf{T_1}$. Note that signs of $\mathbf{b_t}$ and $\mathbf{b_r}$ are chosen such that positive values correspond to a bias towards coding and leading strand, while negative values indicate shifts towards template and lagging strand, respectively.

Transcriptional and replicational signature activities To assess the activity of mutational processes in transcribed versus untranscribed, and early versus late replicating regions, we introduce two additional scalars per signature represented in two vectors $\mathbf{a_t}$ and $\mathbf{a_r} \in {\mathbb{R}_+}^{1\times s}$. Both vectors are stacked and reshaped to match the shape of $\mathbf{T_1}$,

$$\mathbf{A} = \begin{bmatrix} \mathbf{a_t} \cdot \mathbf{a_r} & \mathbf{a_t} \cdot \mathbf{a_r} & \mathbf{a_t} \\ \mathbf{a_t} \cdot \mathbf{a_r} & \mathbf{a_t} \cdot \mathbf{a_r} & \mathbf{a_t} \\ \mathbf{a_r} & \mathbf{a_r} & \mathbf{1} \end{bmatrix}. \tag{2.19}$$

Mutational Composition Quantification of other mutation types requires another $1 \times s$ sized vector $\mathbf{m}$, satisfying the constraint $0 \leq \mathbf{m}_i \leq 1$ for $i = 1, \ldots, s$. In order to include this factor in the tensor factorization, we reshape the vector to $\mathbf{M} \in \mathbb{R}_+^{1\times 1\times 1\times s}$, while $(1 - \mathbf{m})$ is factored into the secondary signature matrix $\mathbf{S}$.

The strand specific-signature tensor

We define the strand-specific signature tensor as

$$\mathbf{T_{strand}} := \mathbf{T_1} \odot \mathbf{B} \odot \mathbf{A} \odot \mathbf{M}, \quad \text{where } \mathbf{T_{strand}} = \mathbb{R}_+^{3\times 3\times p\times s}, \tag{2.20}$$

which therefore subsumes all parameters to describe a mutational process with regard to transcription and replication, and quantifies to what extent the signature is composed of SNVs. To understand this, consider the entry of the signature tensor representative for coding strand mutations, e.g. $(\mathbf{T_{strand}})_{13\cdot\cdot} = \mathbf{b_t} \odot \mathbf{a_t} \odot \mathbf{m} \odot \frac{1}{2}(\mathbf{T_0^{C/G}} + \mathbf{T_0^{C/T}})$, which explicitly states how the low dimensional tensor factors for transcription are broadcasted into the signature tensor.

Signature activities for nucleosomal, epigenetic and clustering states The strand-specific signature tensor $\mathbf{T_{strand}}$ can be considered as the basic building block of the signature tensor, as we instantiate "copies" of $\mathbf{T_{strand}}$ by broadcasting scalar variables for each genomic

state and signature along their respective dimensions. For example, we split SNVs in $t = 3$ nucleosome states (minor in, minor out and linker regions). However, since SNVs may also fall into regions with no nucleosomal occupancy, we distributed mutations across $t + 1 = 4$ states in the corresponding dimension of the mutation count tensor. To fit parameters assessing the activity of each signature along these states, we initialise a matrix $\mathbf{k} \in \mathbb{R}^{(t+1)\times s}$, which can be considered as a composite of a $1 \times s$ constant vector ($\mathbf{k}_{1i} = 1$ for $i = 1, \ldots, s$), and a $t \times s$ matrix of state variables, allowing the model to adjust these parameters with respect to the first row, which corresponds to the non-nucleosomal mutations (baseline). To include these parameters in the factorization, we first introduce singleton dimensions into the strand-specific signature tensor such that $\mathbf{T_{strand}} \in \mathbb{R}_+^{3\times3\times1\times p\times s}$, and reshape $\mathbf{k}$ to match the dimensionality of $\mathbf{T_{strand}}$,

$$\mathbf{k} \in \mathbb{R}_+^{(t+1)\times s} \quad \Rightarrow \quad \mathbf{K} \in \mathbb{R}_+^{1\times1\times(t+1)\times1\times s}. \tag{2.21}$$

Both tensors have now the right shape to enable element wise multiplication with broadcasting

$$\mathbf{T} = \mathbf{T_{strand}} \odot \mathbf{K} \quad \text{where } \mathbf{T} \in \mathbb{R}_+^{3\times3\times(t+1)\times p\times s}. \tag{2.22}$$

We proceed similarly for all remaining genomic properties such as activities along epigenetic domains, and clustering propensities. Generally, to assess l genomic properties, we first introduce l singleton dimensions to the strand-specific signature tensor $\mathbf{T_{strand}}$, instantiate l matrices $\mathbf{k_j} \in \mathbb{R}_+^{(t_j+1)\times s}$ for $j = 1, \ldots, l$ each with t_j states, reshape them appropriately to tensor factors $\mathbf{K_j}$, and broadcast them into the strand specific signature tensor $\mathbf{T_2}$. Here, we introduced new dimensions for epigenetic domains (epi), nucleosomal location (nuc) and clustering propensities (clu), and thus we reshaped the strand specific signature tensor to $\mathbf{T_{strand}} \in \mathbb{R}_+^{3\times3\times1\times1\times1\times p\times s}$, instantiated $\mathbf{k_{epi}} \in \mathbb{R}_+^{16\times s}$, $\mathbf{k_{nuc}} \in \mathbb{R}_+^{4\times s}$ and $\mathbf{k_{clu}} \in \mathbb{R}_+^{2\times s}$ and computed

$$\mathbf{T} = \mathbf{T_{strand}} \odot \mathbf{K_{epi}} \odot \mathbf{K_{nuc}} \odot \mathbf{K_{clu}} \quad \text{where } \mathbf{T} \in \mathbb{R}_+^{3\times3\times16\times4\times2\times p\times s} \tag{2.23}$$

to obtain the final signature tensor $\mathbf{T}$. Note that the parametrisation of nucleosomal, epigenetic and clustering states is generalisable, i.e. it can easily be modified to accommodate for additional or other genetic states (see also A.3.1), and assumes the independence of genomic states.

2.2.3 Error model

The model assumes that the expected values of $\mathbf{C}^{\text{SNV}}$ and $\mathbf{C}^{\text{other}}$ are determined by the inner product of the signature tensor (using the convention that $\times$ is taken over the last dimension of the array on its left – denoting each different signature – and the first dimension of the array on its right) and the exposure matrix and similarly for the non-SNV signature matrix $\mathbf{S}$ and the same exposure matrix $\mathbf{E}$

$$\mathbb{E}[\mathbf{C}^{\text{SNV}}] = \mathbf{T} \times \mathbf{E} \quad \text{and} \quad \mathbb{E}[\mathbf{C}^{\text{other}}] = \underbrace{(\mathbf{S_0} \odot (1 - \mathbf{m}))}_{\mathbf{S}} \times \mathbf{E}. \tag{2.24}$$

To prevent oversegmentation and ensure a robust fit of signatures, we assume that the data follows a negative binomial distribution (Sec. 2.1) with mean $\mathbf{T} \times \mathbf{E}$ and $\mathbf{S} \times \mathbf{E}$, and dispersion τ

$$\mathbf{C}^{\text{SNV}}_{i\ldots n} \sim \text{NB}((\mathbf{T} \times \mathbf{E})_{i\ldots n}, \tau) \quad \text{and} \quad \mathbf{C}^{\text{other}}_{mn} \sim \text{NB}((\mathbf{S} \times \mathbf{E})_{mn}, \tau). \tag{2.25}$$

We use the Tensorflow framework (Sec. 2.1.4) to find the maximum likelihood estimates (MLE) $\hat{\mathbf{T}}$, $\hat{\mathbf{S}}$, $\hat{\mathbf{E}}$ for $\mathbf{T}$, $\mathbf{S}$ and $\mathbf{E}$, respectively, using the parametrisation defined in the previous section. We initialise the parameters of the model with values drawn from a truncated normal distribution, transform them appropriately to meet the requirements of non-negativity, and compute $\hat{\mathbf{T}} \times \hat{\mathbf{E}}$ and $\hat{\mathbf{S}} \times \hat{\mathbf{E}}$ which are fed into the negative binomial likelihood function

$$\mathcal{L}^{\text{SNV}}(\mathbf{C}^{\text{SNV}}_{i\ldots n}; (\mathbf{T} \times \mathbf{E})_{i\ldots n}, \tau) = \prod_{i\ldots n} \frac{\Gamma(\tau + \mathbf{C}^{\text{SNV}}_{i\ldots n})}{\Gamma(\tau)\mathbf{C}^{\text{SNV}}_{i\ldots n}!} \left(\frac{\tau}{\tau + \mathbf{C}^{\text{SNV}}_{i\ldots n}}\right)^{\tau} \left(\frac{(\mathbf{T} \times \mathbf{E})_{i\ldots n}}{\tau + (\mathbf{T} \times \mathbf{E})_{i\ldots n}}\right)^{\mathbf{C}^{\text{SNV}}_{i\ldots n}} \tag{2.26}$$

and

$$\mathcal{L}^{\text{other}}(\mathbf{C}^{\text{other}}_{mn}; (\mathbf{S} \times \mathbf{E})_{mn}, \tau) = \prod_{mn} \frac{\Gamma(\tau + \mathbf{C}^{\text{other}}_{mn})}{\Gamma(\tau)\mathbf{C}^{\text{other}}_{mn}!} \left(\frac{\tau}{\tau + \mathbf{C}^{\text{other}}_{mn}}\right)^{\tau} \left(\frac{(\mathbf{S} \times \mathbf{E})_{mn}}{\tau + (\mathbf{S} \times \mathbf{E})_{mn}}\right)^{\mathbf{C}^{\text{other}}_{mn}}. \tag{2.27}$$

The total log likelihood $\log \mathcal{L}$ is then given by the sum of individual log-likelihoods

$$\log \mathcal{L}(\mathbf{C}^{\text{SNV}}, \mathbf{C}^{\text{other}}; \mathbf{T}, \mathbf{S}, \mathbf{E}, \tau) = \log \mathcal{L}^{\text{SNV}}(\mathbf{C}^{\text{SNV}}_{i\ldots n}; (\mathbf{T} \times \mathbf{E})_{i\ldots n}, \tau) + \log \mathcal{L}^{\text{other}}(\mathbf{C}^{\text{other}}_{mn}; (\mathbf{S} \times \mathbf{E})_{mn}, \tau) \tag{2.28}$$

and thus the optimisation problem is equivalent to maximise the total log likelihood (or equivalently to minimise the negative total log-likelihood)

$$\hat{\mathbf{T}}, \hat{\mathbf{S}}, \hat{\mathbf{E}} = \text{argmin}_{\mathbf{T},\mathbf{S},\mathbf{E}} \left\{ -\log \mathcal{L}(\mathbf{C}^{\text{SNV}}, \mathbf{C}^{\text{other}}; \mathbf{T}, \mathbf{S}, \mathbf{E}, \tau) \right\} \tag{2.29}$$

Moreover, inferring $\hat{\mathbf{T}}, \hat{\mathbf{S}}$, and $\hat{\mathbf{E}}$ enables us to calculate log-likelihood of the MLE

$$\log \hat{\mathcal{L}} = \log \mathcal{L}(\mathbf{C}^{\text{SNV}}, \mathbf{C}^{\text{other}}; \hat{\mathbf{T}}, \hat{\mathbf{S}}, \hat{\mathbf{E}}, \tau). \tag{2.30}$$

2.2.4 Numerical optimisation

Fitting signatures, exposures and tensor factors simply requires to minimise the negative total log-likelihood (Eq. 2.28). However, it is important to note that non-negative matrix factorisation produces "stochastic" solutions, i.e. each decomposition represents a local minimum of the objective function that is used to train the model, and often strongly depends on the initialisation of the parameters. For this reason, it is worthwhile to sample the solution space thoroughly, and to pick the solution which minimised the value of the objective function (see also Sec. A.3.3).

We minimise the negative total log-likelihood using an ADAM Grad optimiser with an exponentially decreasing learning rate and starting value of 0.1. The number of epochs to train the model depends on the size of the dataset. Generally, the model should be trained long enough to achieve convergence of the objective function, i.e. the gain of log-likelihood between consecutive epochs should tend towards zero (see also Sec. A.3.3).

2.2.5 Model selection

Choosing an appropriate column rank prior to the decomposition is crucial to determine mutational signatures accurately. The algorithm fails to detect mutational processes if the chosen decomposition rank is too small, and may infer artefact signatures if it is too large. Alexandrov et al. (2013b) determine the rank by selecting s such that the reconstruction error of **SE** is low and the reproducibility of solutions is large. In our framework, this problem is slightly more complicated, because in addition to the right column rank, we have to identify the appropriate dispersion τ for the data as well. For this reason, we decided to fix τ prior to the analysis, and only compare models of different decomposition rank across the ones which were fitted using the same dispersion.

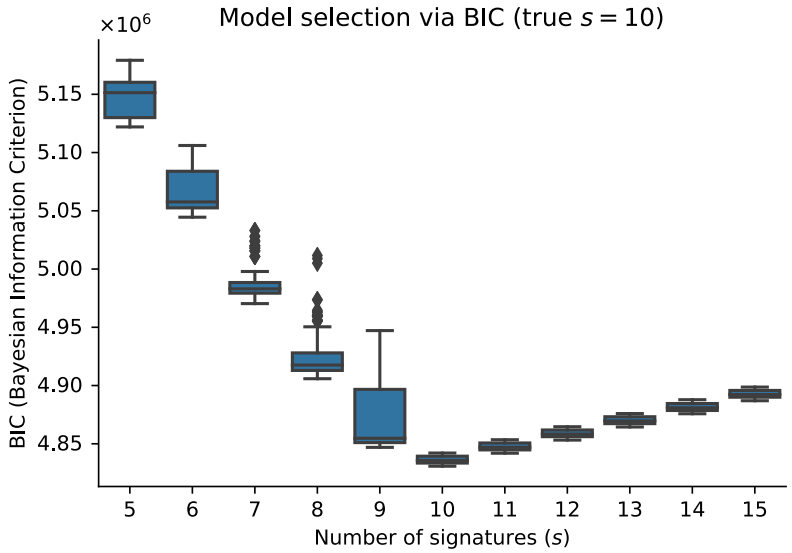


Fig. 2.7 **Choosing the appropriate column rank of the signature decomposition using the BIC.** In this simulation experiment, we simulated mutation counts using 10 mutational signatures and random exposures of 100 samples, and subsequently applied TensorSignatures decompositions with the dispersion used to simulate the data. Applying the BIC to these extractions accurately identifies the true number of signatures as indicated by the kink of the BIC curve.

To select the appropriate number of signatures s for a model with dispersion τ, we compute for each decomposition rank the Bayesian Information Criterion (BIC, Schwarz (1978))

$$\mathrm{BIC}_\tau(s) = \log(n) \cdot k(s) - 2 \cdot \log \hat{\mathcal{L}}, \tag{2.31}$$

where n is the number of observations (total number of counts in $\mathbf{C}^{\mathrm{SNV}}$ and $\mathbf{C}^{\mathrm{other}}$), k represents number of parameters in the model (which depends on the rank s), and $\log \bar{\mathcal{L}}$ is the log-likelihood of the MLE. The BIC tries to find a trade-off between the log-likelihood and the number of parameters in the model; chosen is the rank which minimises the BIC. In our simulation experiments, we found that the BIC accurately determines the appropriate column rank given the correct dispersion of the simulated data (Fig. 2.7).

2.2.6 Bootstrap Confidence Intervals

To compute bootstrap confidence intervals (CIs) for inferred parameters, we randomly select $\frac{2}{3}$ of the samples in the dataset, initialise the model with the MLE for $\hat{\mathbf{T}}$ and $\hat{\mathbf{S}}$ while randomly perturbing the 10 % of their estimates, and subsequently refit $\hat{\mathbf{T}}$, $\hat{\mathbf{S}}$ and $\hat{\mathbf{E}}$ to the subset of

samples. Initialising the parameters with the MLE results from computational constraints, as this step needs to be repeated for 300 - 500 times to obtain a representative distributions of the parameter space. Next, we match refitted signatures to the MLE reference by computing pairwise cosine distances, and accept bootstrap samples if the total variation distance between the bootstrap candidate and the reference is smaller than 0.2. Finally, we compute 5 % and 95 % percentiles on accepted bootstrap samples to indicate the CIs of our inference.

2.3 Assessment of TensorSignatures

2.3.1 Simulation Studies

To learn more about the conditions that affect the ability to extract tensor signatures, we simulated data under different scenarios, and compared obtained solutions with the ground truth. If not otherwise specified, we selected a pre-described number of COSMIC signatures and multiplied them with randomly generated tensor factors (transcription biases etc.), and exposures from 10 to 1,000 genomes. We chose the simulation parameters such that the average number of mutations per genome ranged between 50 and 1,000, and created different realisations of the data by perturbing the resulting counts ten times with negative binomial noise ($\tau = 50$). We performed each tensor factorization 10 times on each of these ten realisations and selected for each realisation the solution in which the negative log-likelihood was minimised. All models were trained for 50,000 epochs using an ADAMgrad optimiser and a constant learning rate of 0.1. Shown figures in this section always report means and standard deviations.

Accuracy of signature and exposure inference

Most importantly, mutational signature analysis should accurately decipher the mutational patterns underlying different mutational processes, and equivalently determine their mutational loads in each genome of the dataset. To assess how the number of mutations per sample (m) and the size of the dataset (n) affect the quality of signature extraction in a cancer catalog simulated from five mutational processes, we compared extracted and true spectra by computing cosine distances (indicated as signature recognition, i.e. $1 - cosinedist(\text{true}, \text{inferred})$), and the difference between true and deciphered exposures (Fig. 2.8a and b respectively). This analysis revealed that either 100 genomes with 100 mutations, or 10 samples with 1,000 mutations each suffice to accurately determine signature spectra with a signature recognition of 0.9 or higher (Fig. 2.8a). A similar cutoff may be chosen to accurately determine cor-

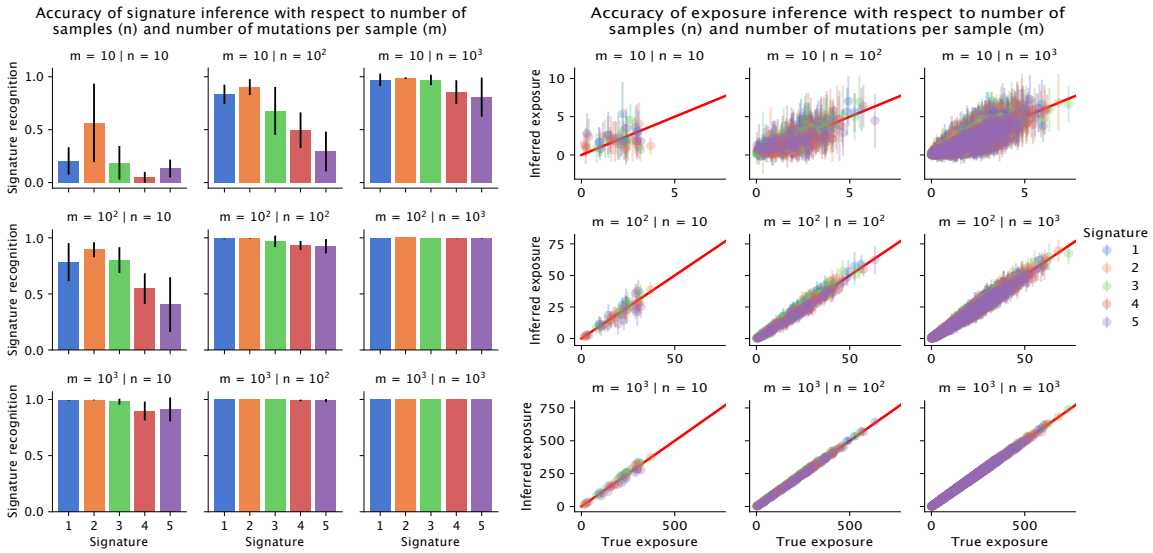


Fig. 2.8 **Accuracy of tensor signature inference. a**. Accuracy of signature inference with respect to the number of samples (n) and the number of mutations per sample (m) in the simulated dataset. Signature recognition is defined as 1 minus cosine distance of the inferred and true signature. **b**. Accuracy of exposure inference with respect to the number of samples (n) and the number of mutations per sample (m) in the simulated dataset.

responding exposures, whose relative errors decrease with increasing mutation and sample numbers (Fig. 2.8b).

Relative errors of tensorfactors quickly decrease as the number of mutations per sample and the dataset size increases

Likewise, we tested the algorithm's ability to infer tensor factors quantifying the genomic properties from mutational signatures in simulated datasets with different number of mutations and sample sizes (Fig. 2.9). To ensure that these tests closely resemble real-world scenarios, we simulated the number genomic dimensions and states similarly to the analysis presented in this work, i.e. three genomic dimensions resembling epigenetic, nucleosomal, and clustered mutations with 16, 4 and 2 states, respectively. The results of this analysis suggest that relative errors of all tensor factors quickly decrease as the size of the cancer catalogue and the number mutations per sample increase (Fig. 2.9).

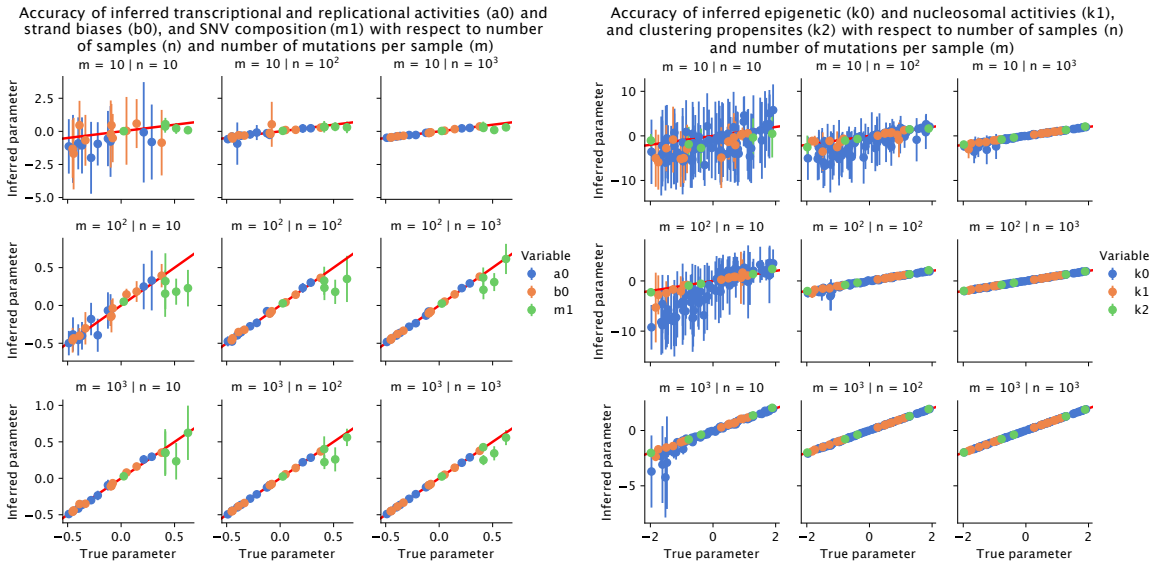


Fig. 2.9 **Relative errors of tensor factors diminish with increasing numbers of genomes and mutation numbers. a**. Accuracy of inferred transcriptional and replicational activities (a_0) and strand biases (b_0), and SNV composition (m_1) with respect to the number of samples (n), and the number of mutations per sample (m) in the simulated dataset. **b**. Accuracy of inferred epigenetic (k_0) and nucleosomal activities (k_1), and clustering propensites (k_2) with respect to the number of samples (n) and the number of mutations per sample (m) in the simulated dataset.

The number of extractable tensor signatures depends on the size of the dataset and the average number of mutations per sample

Next, we tested how many tensor signatures the algorithm reliably detects given a fixed number of mutations and a varying number of samples, and conversely, given a fixed number of genomes and a varying number of mutations per sample (Fig. 2.10). The number of mutations per sample strongly influences the accuracy of signature detection. While it is theoretically possible to detect 40 signatures with high accuracy using a dataset comprising only 100 samples with 10,000 mutations, it requires at least 2,000 genomes to achieve a similar quality of signature recognition with genomes containing 1,000 mutations (Fig. 2.10). These results are underpinned by the complementary experiment shown in Fig. 2.10. Although the signature recognition improves considerably from additional samples in simulation experiments with 100 and 1,000 mutation per samples, respectively, the accuracy of signature detection hardly improves from additional samples once these contain 10,000 mutations on average.

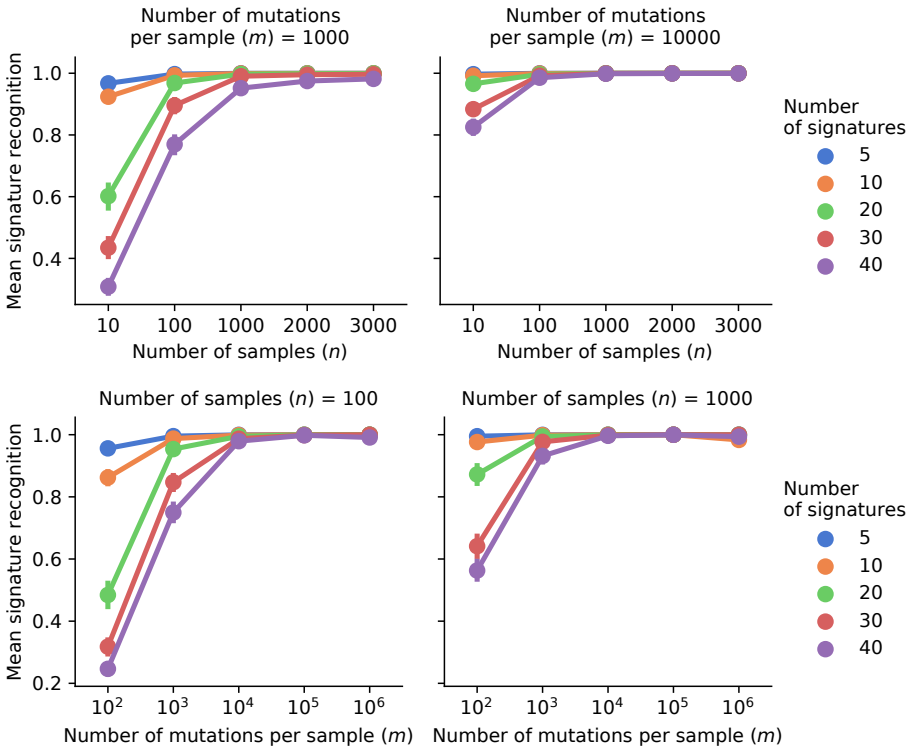


Fig. 2.10 **Signature recognition benefits from larger datasets and mutation numbers per sample.** Accuracy of signature recognition at different ranks with respect to sample size (n) and number of mutations (m).

2.3.2 Comparison of TensorSignatures to conventional NMF methods

TensorSignatures addresses many unresolved issues in the field of mutational signature analysis. For example, it allows to natively quantify the activity of mutational processes in various genomic regions and links other mutation types to SNV spectra. To generate similar results, previous approaches used to regress out the activity of mutational processes within specific genomic domains, or made use of post-hoc "posterior" calculations. Another common practice is to correlate the exposures of independent NMFs to associate the spectra from other mutation types to corresponding SNV signatures. In this section, we compare the algorithm to these aforementioned approaches.

Extracting the genomic properties of mutational signatures by regression

We tried to quantify the genomic properties of mutational signatures using a less principled approach by simulating a mutation count tensor ($\mathbf{C}^{\text{SNV}} \in \mathbb{N}_0^{3\times3\times16\times4\times2\times p\times n}$). To recover mutational signatures and corresponding sample exposures, we factorised the marginalised

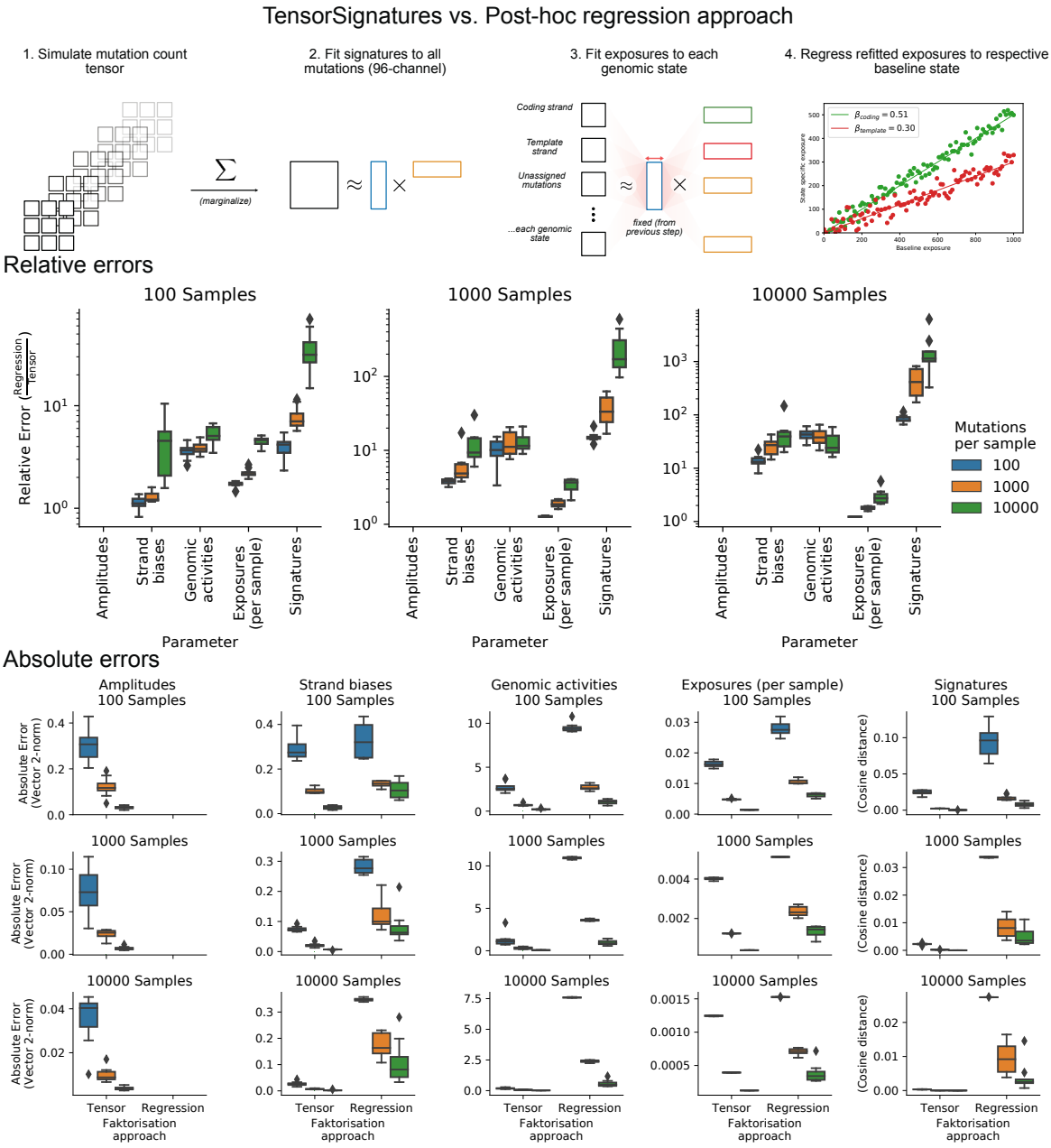


Fig. 2.11 **TensorSignatures determines the properties of mutational signatures more accurately in comparison to post-hoc regression approaches.** Upper panel: Schematic depiction of a simulation approach which recovers signature activities similar to TensorSignatures by regressing their prevalence in different genomic dimensions post-hoc. Lower two panels depict relative and absolute errors respectively.

(summed) single base substitution count tensor on the 96-trinucleotide and sample dimension ($\mathbf{C}^{\text{marg}} \in \mathbb{N}_0^{p \times n}$). To determine strand biases and signature activities across genomic states, we fixed extracted spectra, and refitted the exposures to the count matrices containing the

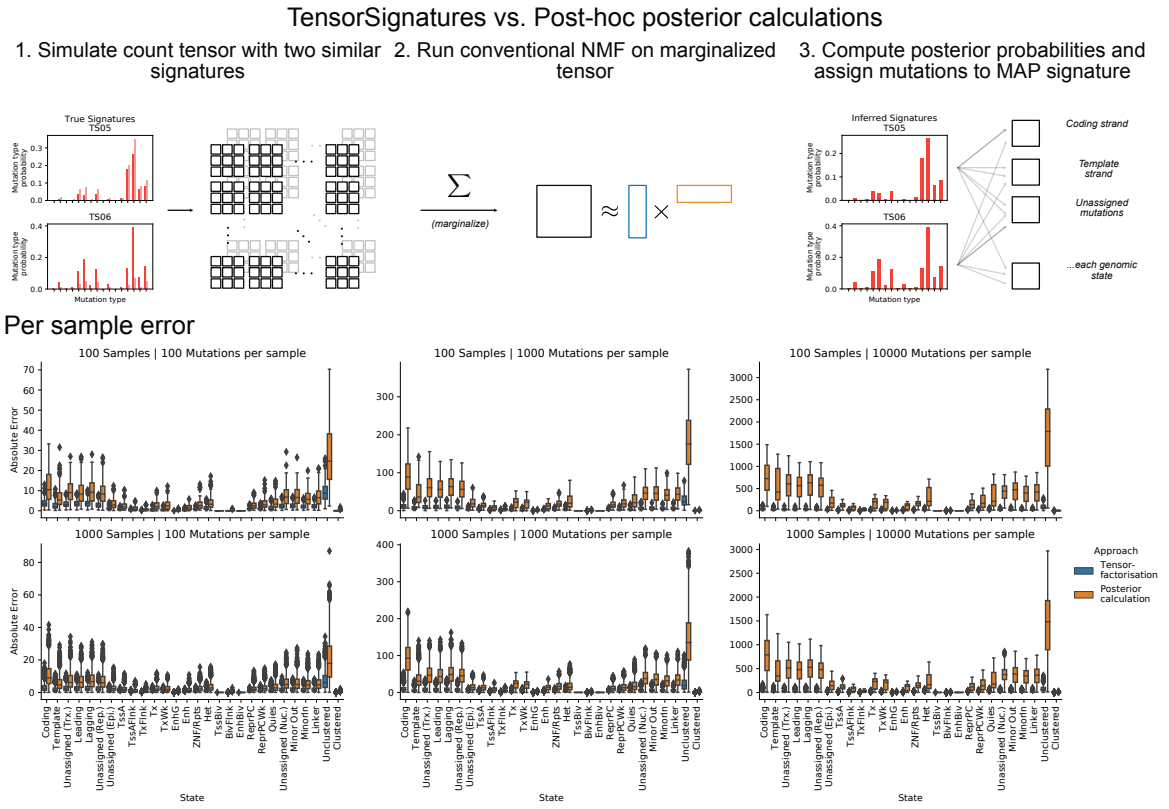


Fig. 2.12 **TensorSignatures assigns mutations more accurately to genomic compartments in comparison to post-hoc posterior calculations.** Upper panel: Schematic depiction of the simulation approach which uses a maximum a posteriori (MAP) estimation to assign mutations to their respective mutational signature as described by Morganella et al. Lower panel: Per sample absolute errors of the tensor factorisation and the aforementioned strategy.

single base substitutions of a specific state only (e.g. template strand mutations, TssA). To obtain a scalar parameter similar to TensorSignatures' tensor factors, we regressed state specific exposures to respective baseline exposures (e.g. exposures of template strand mutations against exposures of unassigned mutations), and compared obtained regression coefficients with the equivalent parameter of the tensor factorisation and the ground truth. To assess the error, we computed the vector 2-norm and cosine similarity for strand biases, genomic activities and exposures, and signature spectra, respectively. We performed this experiment for datasets with sizes (100; 1,000; 10,000) and different numbers of mutations per sample (100; 1,000; 10,000). Note that this approach fails to recover signature activities in untranscribed/transcribed and early/late replicating regions (indicated as "Amplitudes" in the following Fig. 2.11).

Our simulations revealed increasing relative errors for all assessed parameters as sample size and mutation loads increase (Fig. 2.11, middle panel). To understand this, consider that only TensorSignatures may leverage the additional information encoded in the tensor representation of larger datasets to improve the estimates of signature defining properties such as strand biases and genomic activities. Although it is possible to find reasonable parameter estimates by fitting signature spectra first and subsequently regressing out the effect of genomic determinants, absolute errors are always larger at similar samples sizes and mutation loads (Fig. 2.11, lower panel).

Assigning single base substitutions to their source signature with maximum a posteriori approaches

To compare the ability to assign mutations to their respective source signature, we designed a simulation experiment in which we used two very similar signatures (TS05 and TS06) to simulate a mutation count tensor. We then applied conventional NMF on the marginalised (summed) count tensor and determined the maximum a posteriori (MAP) signature for each trinucleotide context in each simulated sample as described in (Morganella et al., 2016).

Absolute errors (vector 2-norm) of post-hoc assigned single base substitutions increase as the number of mutations per sample get larger, while the predictions of the equivalent tensor factorisation become more accurate. This is expected as the post-hoc signature posterior probability is only conditioned on the mutation type and the sample exposure. Furthermore, shown results are likely to underestimate errors as our simulations and inferences were performed using only two signatures, and thus correct signature assignment is likely to happen by chance (Fig. 2.12).

Stability of solutions

Another challenge in mutational signature analysis is the problem of unambiguously associating other variant types to respective single base substitution spectra. Common practice is to perform independent NMFs on each variant type, and to subsequently match mutation subtype specific signatures to the SBS correlate by assessing exposures. In contrast, TensorSignatures decomposes SNV and other mutation type counts simultaneously, thus circumventing the problem of post-hoc associating different mutation types, and delivering a more robust signature inference by pooling the evidence from the entire mutational imprint.

To illustrate this, we ran independent NMFs on SNV and other mutation count matrices of the PCAWG dataset. To match resulting mutational spectra, we computed the correlation coefficients of exposures and paired highest correlating SNV and other mutation type

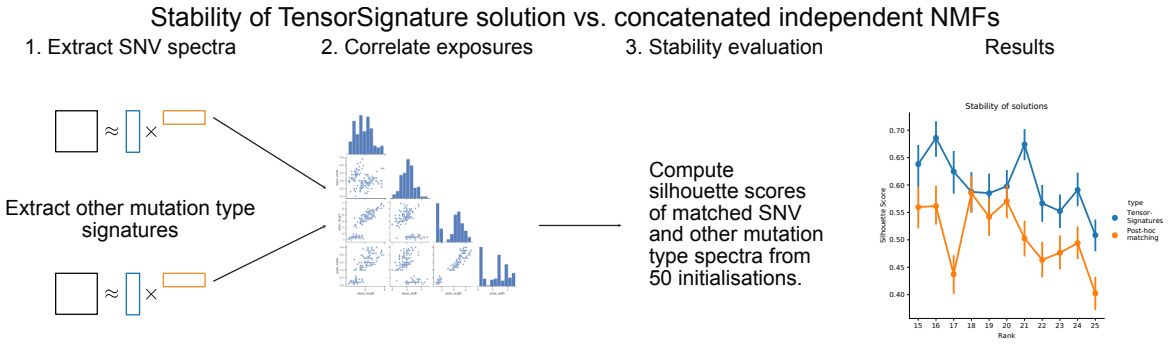


Fig. 2.13 **TensorSignatures assigns other mutation types more confidently to associated SNV spectra.** Left panel: Matching SNV and other mutation type spectra (from independent NMFs) by correlating their exposures. Right panel: Signature stability across several ranks using TensorSignatures and the signature matching approach.

signatures. We repeated these steps 50 times to obtain a set of 50 initialisations of paired mutational signatures (SNV + other mutation types), and compared the stability of these solutions with TensorSignature decompositions by computing the silhouette scores across several ranks (Fig. 2.13). Our results indicate a higher stability of TensorSignatures solutions across all tested ranks implying that the tensor framework more consistently reproduced SNV and their accompanying other mutation type spectra.

TensorSignatures in context of other tools

There is a wide range of software packages that aim to extract mutational signatures. In this section, I will review a selection of tools which differ in their methodology and application, and discuss them while relating them to TensorSignatures (a brief comparison table is provided in Tab. B.6).

SigProfiler SigProfiler is perhaps the most commonly used tool in mutational signature analysis, and has been used in landmark studies such as PCAWG (Alexandrov et al., 2020). The software comprises of a set of utilities including SigProfiler MatrixGenerator, SigProfiler Extractor, SigProfiler Plotting (Bergstrom et al., 2019), and SigProfiler Simulator (Bergstrom et al., 2020) which enable convenient data preparation, signature extraction and visualisation, as well as the simulation of synthetic cancer genomes, respectively. In comparison to SigProfiler, TensorSignatures allows the concurrent extraction of mutational signature across all variant types, and accounts for a wide range of different genomic factors.

Sparse Signatures SparseSignatures puts particular emphasis on avoiding signature overfitting by applying a LASSO penalty on signatures to favor the extraction of "well-differentiated"

spectra, and by taking into account the standard replication error (Lal et al., 2020). Finally, it implements cross-validation for model selection, which according to the authors, more reliably detects the appropriate number of signatures. In contrast, TensorSignatures addresses overfitting by employing a robust noise model based on an overdispersed negative binomial distribution to model mutation count data, and uses the conservative BIC estimator to identify the most likely number of signatures.

deconstructSigs DeconstructSigs focuses on fitting the exposures of a pre-defined set of signatures to new datasets, which is useful when only a few samples are available, and therefore the denovo extraction of mutational signatures is infeasible (Rosenthal et al., 2016). The software uses a multiple linear regression model with the caveat that any coefficient must be greater than 0, as negative contributions make no biological sense. TensorSignatures provides a similar feature, i.e. it enables to refit a set of tensor signatures to novel samples if the user runs the tool using the `refit` option (Sec. Appendix A).

EMu EMu is a conceptually interesting tool as it deciphers the activity of mutational signatures both genome-wide and within local regions of the genome (Fischer et al., 2013). However, in comparison to TensorSignatures, which relies on state assignments to SNVs, EMu divides each genome into non-overlapping 1Mb regions and extracts the activity of each signature globally as well as locally. This approach enabled the authors to associate certain mutational processes to phenomena like kataegis and epigenetic features such as chromatin modifications. From a methodological point of view, EMu employs an expectation-maximization algorithm to conduct the inference and uses the Bayesian information criterion to perform model selection.

Mutalisk Mutalisk is a web-based tool that identifies up to seven active signatures from a set of reference signatures using linear regression, and assesses uploaded VCF files with respect to kataegis, transcriptional strand bias, DNA replication timing, GC content and histone modifications (Lee et al., 2018). The tool computes correlation coefficients between aforementioned features and the provided list of somatic mutations and displays these conveniently on the website. In contrast, TensorSignatures deconvolves mutational signatures while accounting for such genomic factors. This enables TensorSignatures to associate mutational processes with genomic features rather than only certain mutation types.

2.4 TensorSignatures in the cloud

Running TensorSignatures on larger datasets requires a GPU with CUDA support to make use of hardware accelerated tensor computations. Additionally, identifying an appropriate solution requires the user to run the analysis for different decomposition ranks and multiple initialisations, suggesting to run Tensorsignatures ideally in parallel with multiple GPUs. These entry requirements may be prohibitive for some users, which is why I decided to built a web application enabling the community to run TensorSignatures online, thus providing a low entry barrier for new users to try out the software. In this chapter, I outline how modern web and cloud technologies may be employed to deploy research pipelines, and demonstrate this by explaining the TensorSingatures web application stack.

2.4.1 Deploying research pipelines using microservices

Cloud computing platforms encourage to break down larger applications to atomic, independently running components, called microservices that interact and communicate with each another, thus providing the accustomed experience from a single application. Each individual service focuses on a single task, while the ensemble communicates through HTTP calls (request/response). The big advantage of microservices are apparent: decoupled services can be developed, deployed, and scaled on their own, allowing them to be upgraded quickly and on demand. In larger applications, clear separation between services enables to easily split responsibilities between many developers. For example, front-end developers can focus on the client facing website, while treating the backend as a blackbox, because proper communication between services is ensured via their application program interfaces (APIs). Because each service works independent from the rest of the application, microservices tend to have smaller code bases that are easier to test, refactor and scale. Clear separation also ensures that errors are localised, such that breaking a single service does not automatically affect the whole application.

However, the microservice architecture has also many drawbacks. First of all, deciding to split an application into many pieces is difficult, as it requires to split apart whole components into independently working units, which is in most cases much harder then to refactor parts to separate modules. For example, it is quite difficult to isolate stateful layers like databases or task queues which require some form of data persistence, such that their state is not shared or duplicated. The microservice pattern also increases the network complexity of the application, since operations that used to be handled within function calls of a single process, now often require network calls to different services, requiring to carefully coordinate APIs

to properly work with each other. Also, it is highly recommended to write more complex integration tests when developing applications with the microservice architecture to ensure that independent parts work well together. Finally, with multiple services, complexity shifts from the codebase to the infrastructure, which might be more costly and difficult to maintain.

2.4.2 Building modern web appplications with Docker and Kubernetes

The reason why microservice patterns became recently more popular are the two core technologies Docker and Kubernetes (k8s). Docker packages applications and their dependencies in containers that can run on any Linux server, which helps to provide portability, enabling microservices to be run in various locations. The framework uses native resource isolation features of linux kernels, which avoid much of the overhead that would otherwise be consumed by virtual machines (VMs), making Docker containers lightweight such that a single server can run many containers at once. On the other hand, k8s allows to automate scheduling, configuration, supervision and failure handling of (Docker packaged) microservices. In this way, k8s shifts the responsibility of deploying an application from the operations[3] team to the developers, enabling them to update and release their work themselves. K8s also takes care of monitoring and rescheduling applications in case of hardware failures, allowing system administrators to shift their focus from individual apps (often many hundreds in large data centres) to ensure that k8s is up and running.

Docker container package microservices

The biggest problem in deploying applications with smaller components is that developers have to deal with differences in the environments they run their microservices on. A multitude of factors contribute to these discrepancies, ranging from different hardware setups of development and production machines to operating systems and available libraries. One solution to this problem are VMs and container technologies, which dedicate resources of a host machine to provide isolated environments, thus allowing to run the same operating system, libraries and configuration during development and production. However, since VMs produce large overheads and require more configuration, linux container prevailed, mainly because they only consume the resources the application requires, thus allowing to run a larger number of microservices with the same resources.

Although container technologies existed for a long time, they only became widely adopted with the rise of the Docker platform, which made the container system easily portable and

[3] Operations refers to the set of processes and services that are administered by an IT department within a larger organisation or business.

simplified the process of packaging up whole applications including their dependencies. However, Docker is more than that, as it also provides a platform for packaging, distributing and running applications, making it possible to transfer packages to a central repository, from which it can be downloaded to any machine with Docker and subsequently be executed. To understand this better, it is necessary to define the most important Docker concepts, namely *Images*, *Registries* and *Containers* (Fig. 2.14). A Docker Image is defined by the instructions of a Docker file that tells Docker exactly what libraries to install, and what application files should be copied and executed when the image is run. Importantly, Images are immutable meaning that they represent an unmodifiable snapshot of an application at a given point in time, ensuring that the programme will work always in a predictable way. The Docker Registry is an online repository for Docker Images facilitating the exchange of images between different computers. Upon building a Docker Image, it can either be directly run on the development machine, or pushed (uploaded) to a registry, from which another machine can pull (download) the image and run it as well. A Docker Container is created from an Image and represents the actual process that runs on the host machine. This process is completely isolated from any other process of the host, and only has access to limited amounts of resources.

Kubernetes orchestrates Docker containers

K8s was developed at Google and released as an open-source project in summer 2014. The software is a successor of Google's proprietary systems Borg and Omega, that were used to ship many of their containerised apps such as Gmail. In essence, k8s abstracts the hardware infra structure of large data centres and exposes it as single computational resource, allowing to develop applications with many components without having to understand the server architecture under the hood. Upon deploying an multi-component application, k8s selects a server for each microservice, and sets up the infrastructure to enable communication between each part of the app. This makes k8s especially valuable to cloud providers, because it allows them to provide a simple interface to their customers, while their system administrators do not have to worry about the manifold of applications that runs on their hardware. One such provider is the EMBL-EBI embassy[4], which provides the k8s infrastructure to deploy the TensorSignatureOnline application.

K8s can be understood as an orchestrator for microservices that make up the building blocks of larger applications. To illustrate this, consider a simple application with two components: a database which stores and retrieves data, and a front-end interface, allowing users to interact with the former. In this example, the database and the front-end represent

[4]https://www.embassycloud.org

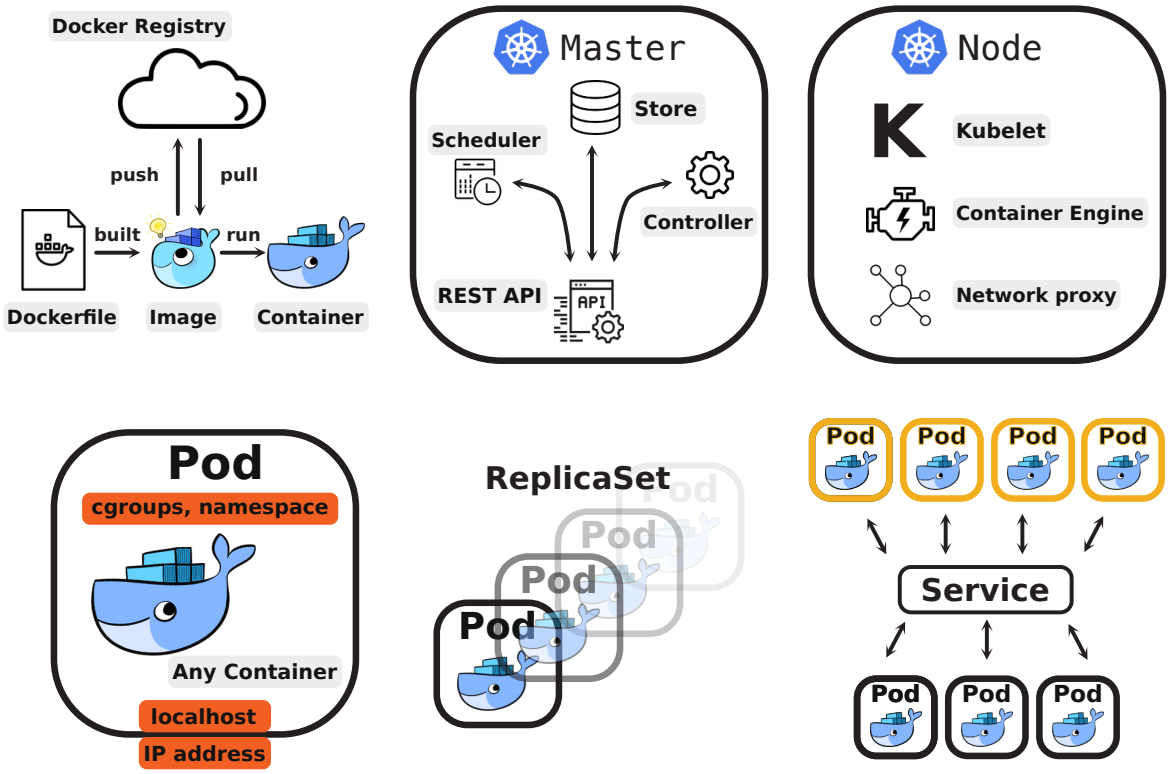


Fig. 2.14 **Docker and Kubernetes.** **a**. Docker enables to run linux based microservices within containers, and provides a distribution platform such that images can be downloaded on any machine running Docker. **b**. A k8s master node accepts declarative instruction via a REST API, saves these and in an internal cluster store and contains a scheduler and controller unit. **c**. A k8s node communicates with the master via its kubelet, pulls and starts Docker images and containers respectively using its container engine, and enables communication between pods through a network proxy. **d**. A pod encapsulates a running Docker container. **e**. Pods are deployed as a part of a ReplicaSet or Deployment, which not only specify which Docker image to run, but also how many copies of such pods should be initiated, as well as update instructions etc. **f**. To reliably connect different Pods, services are interposed ensuring that requests to appropriately transmitted to respective pods.

microservices, which each run in dedicated containers, while k8s orchestrates both services to ensure that both interact properly with each other to provide a useful application.

Kubernetes master and nodes A k8s cluster consists of one or more *masters* and several *nodes*. The masters assign microservices to the nodes, monitor them and implement changes upon events. In order to achieve this, masters contain four major components: an API server, the cluster store, the controller manager and the scheduler (Fig. 2.14b). The API server simply represents a RESTful API, that accepts YAML or JSON manifests specifying instructions for microservices. If these pass the validation step of the API server, they get

passed to the cluster store which can be considered as the memory of an k8s cluster, as it stores the clusters' configuration. While the controller manager implements a few functions, the scheduler performs resource management and assigns workloads to nodes. In summary, masters run all processes necessary to control and schedule workloads, and enable users to interact with the k8s cluster.

In contrast, nodes run microservices, report back to the masters, and watch out for work assignments. They comprise three units: the kubelet, the container runtime, and the kube-proxy (Fig. 2.14c). The kubelet is probably the most important piece, as it registers the host to the k8s cluster and surveils the API server for new work assignments upon which it triggers requested steps and reports back to the master about the outcome. If the kubelet, for example, fails to run a particular task, it reports back the failure and the master decides about subsequent actions. The container runtime receives instruction from the kubelet and executes all tasks necessary to manage containers, i.e. pulling images, or starting and stopping containers. By default the container runtime uses Docker to perform all container related tasks, although it is possible to run other container runtime software as long as these satisfy the the k8s container runtime interface. The last important part is the kube-proxy, which handles all network related tasks such as assigning unique internet protocol (IP) addresses to each container of the node, and enables load-balancing.

Declarative object management and desired state principle Central to k8s are the concepts of declarative object management and desired state principle. To appreciate them, consider the deployment of k8s applications: The manifest file defines the desired state of the application by declaring parameters such as the specific Docker image, number of replicas, and update rules. After posting the file to the API server, k8s inspects and records the manifest in the cluster store as part of the cluster's overall desired state. Then, the nodes of the cluster take over, which pull respective images and build the desired network. To ensure that the cluster does not deviate from its desired state, k8s sets up watch loops that constantly surveil the current state of the cluster, and trigger appropriate actions if the cluster varies from it. Importantly, this process is handled in a declarative manner, i.e. rather than specifying a long list of instructions leading to desired outcome, we simply declare the desired state, and k8s takes care of the implementation. This paradigm has various advantages, for example, it enables self-healing, automated scaling, version control and self-documentation.

Kubernetes Pods represent the atomic unit of deployment Rather than directly running docker container on k8s nodes, Kubernetes packages containerised apps into so-called pods which represent the atomic unit of deployment (Fig. 2.14d). A pod can be conceptualised

as a protected environment or sandbox to run one or multiple containers. It establishes a ring-fenced area of the host operating system to build a network stack, creates kernel namespaces and exposes them to the inside container. For example, if multiple container run within a single pod, they all share the same environment, including memory, volumes, network stack and IP. Since pods are the smallest unit of deployment, they allow scaling simply by creating several replicas of them (this is known as "horizontal scaling"). Also, they only have two states, they are either up and running, or they fail, in which case k8s discards them and spins up a novel pod to replace the broken instance.

Pods are deployed as a part of ReplicaSets or Deployments Pods are normally deployed as a part of *ReplicaSets* or *Deployments*. The former simply takes the template of a pod and deploys it for the desired number of replicas, while also instantiating background watch loops to ensure that the number of active pods on the cluster does not deviate from the desired state. The latter represents an higher level abstraction that encapsulate ReplicaSets, thus not only providing their features, but also enable to define update and rollback models. ReplicaSets and Deployments, like any other resources in k8s, are instantiated by defining a declarative manifest in YAML or JSON format and are posted to the API server of a k8s master node.

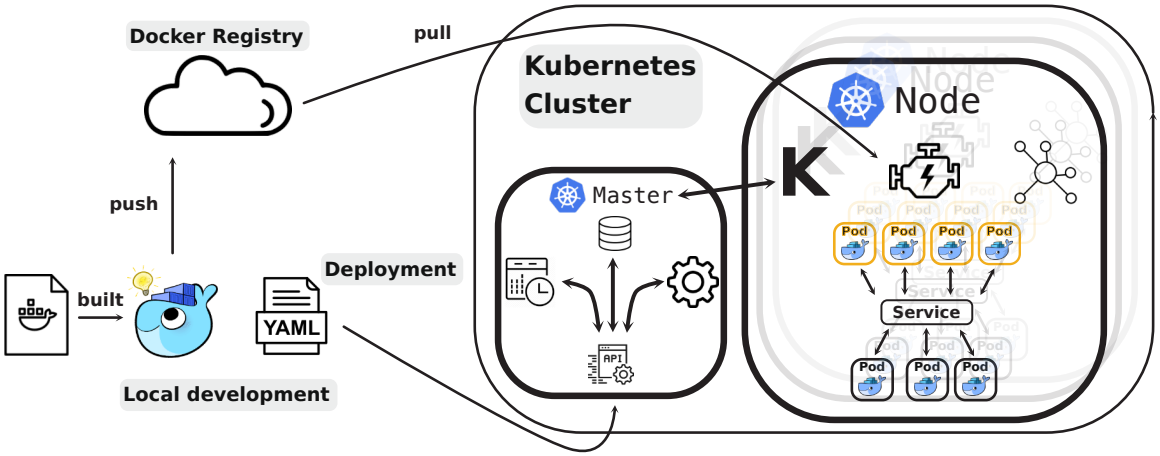


Fig. 2.15 **Deploying applications with Docker and Kubernetes.** To deploy an application with Docker and Kubernetes, microservices need to be developed and packaged locally. These are then pushed to the Docker registry, while at the same time, declarative instructions are posted to the Kubernetes cluster. The Kubernetes cluster then takes over and instructs its nodes to implement the application by pulling and installing respective microservices.

Services enable communication between different pods The ephemeral nature of pods makes it difficult to reach them through the cluster internal network, e.g. upon failure, scaling

or updating the application, k8s instantiates new pods, each with a novel and unique IP address. Since the microservice architecture heavily relies on the communication between different pods, k8s provides so-called *Services* to ensure reliable networking endpoints for a logical set of pods that perform a similar function (Fig. 2.14f). In this way, services represent a persistent abstraction layer over pod(s), while enabling service discovery and load balancing.

Persistent Volumes and Claims Although pods share resources such as CPU and RAM with other pods running on the same host, they do not share disk storage, as each individual container within a pod has its own isolated filesystem. As a result, pods only have access to the files that were added to the image at build time, as well as to the files written by the image to the pod internal disk at runtime, but not to any other external file. However, in many scenarios it is desired to preserve data, for example to ensure that a pod starts again at the exact same state after its life-cycle came to an end, or in data processing pipelines when different pods need access to a particular file. K8s enables this by providing persistent volumes which define a storage volume that is independent of the normal pod-lifecycle. Pods need to be issued with a persistent volume claim (PVC) to access and write a persistent volume.

Putting it all together Deploying an application using the microservice pattern with Docker and Kubernetes involves in principle two steps (Fig. 2.15). First, developing and packaging all required microservices with Docker locally, and specifying corresponding *Deployment* files in form of YAML manifests, which tell the cluster how each service should look like (what image to use, which ports to expose etc.), and how many replicas to create of each component. Second, building the images, pushing them to the Docker registry, and posting the Deployment files to the cluster. From here Kubernetes takes over, and takes care of implementing the application by pulling all required microservices and setting up an appropriate network.

To accomplish this, the master node validates incoming deployments and stores the desired state of the application within its cluster store, whereupon the scheduler assigns the workload to one of the listening nodes. A node receives the instructions from the master via its kubelet, which subsequently instructs its container engine to pull requested images from the Docker registry. Then the container engine launches respective containers within dedicated pods, while the node's network proxy puts in place the required network between microservices. Finally, the kubelet monitors the deployment and reports back failures to the master (Fig. 2.15).

2.4.3 The TensorSignaturesOnline web application

With TensorSignaturesOnline I set out to create a web application that enables users to

1. inspect the results from our mutational signature analysis on the PCAWG dataset,
2. and allowing them to create user accounts that gives them access to a protected area at which they can upload their own VCF data, and fit the exposures to the set of PCAWG tensor signatures to their samples.

To achieve these goals, I built a web application based on the Python web server framework Flask, a PostgreSQL database, and a Redis task queue. In the following, I describe the structure and features of the application, and explain the thoughts that went into designing the stack.

A Flask framework serves the back- and frontend of TensorSignaturesOnline

Flask is an extensible python web framework which serves in TensorSignaturesOnline both as back- and frontend. The backend implements all routes (views) of the website, handles user authorisation, and dispatches jobs to a Redis task queue. To enable this, the backend tracks users, tasks and datasets within a PostgreSQL database, whose structure I describe in the following section. Note, that the Flask pod has access to a persistent volume which allows the app to receive and store user uploaded VCF data (Fig. 2.16).

Views of TensorSignaturesOnline In order to render the client viewing webpage, we used Flasks template engine, which contains static HTML markup as well as placeholders for dynamic data. To enable proper display of the website across different devices, we employed the Bootstrap CSS framework to make the sites responsive. The pages available in the application can be divided into those which are available to everyone, and some which can only be accessed after a registration to the service. Among the routes which are available to unregistered users is the welcome page (route: `/`, Fig: 2.17a); a overview page of all tensor signatures (route: `/signatures`, Fig. 2.17b); a detailed page for each tensor signature, including their spectra, genomic properties and a list of samples with largest exposures (route: `/signature/<signature_id>`, Fig. 2.17c); an overview page for each cancer type (route: `/cancer/<cancer-type>`, Fig. 2.17d); and a view for each sample encompassing the spectra and a detailed breakdown of each active signature (route: `sample/<dataset_id>/<sample_id>`, Fig. 2.17e). After logging into the application, users obtain access to a dashboard (Fig. 2.17f) allowing them to create datasets, i.e. a

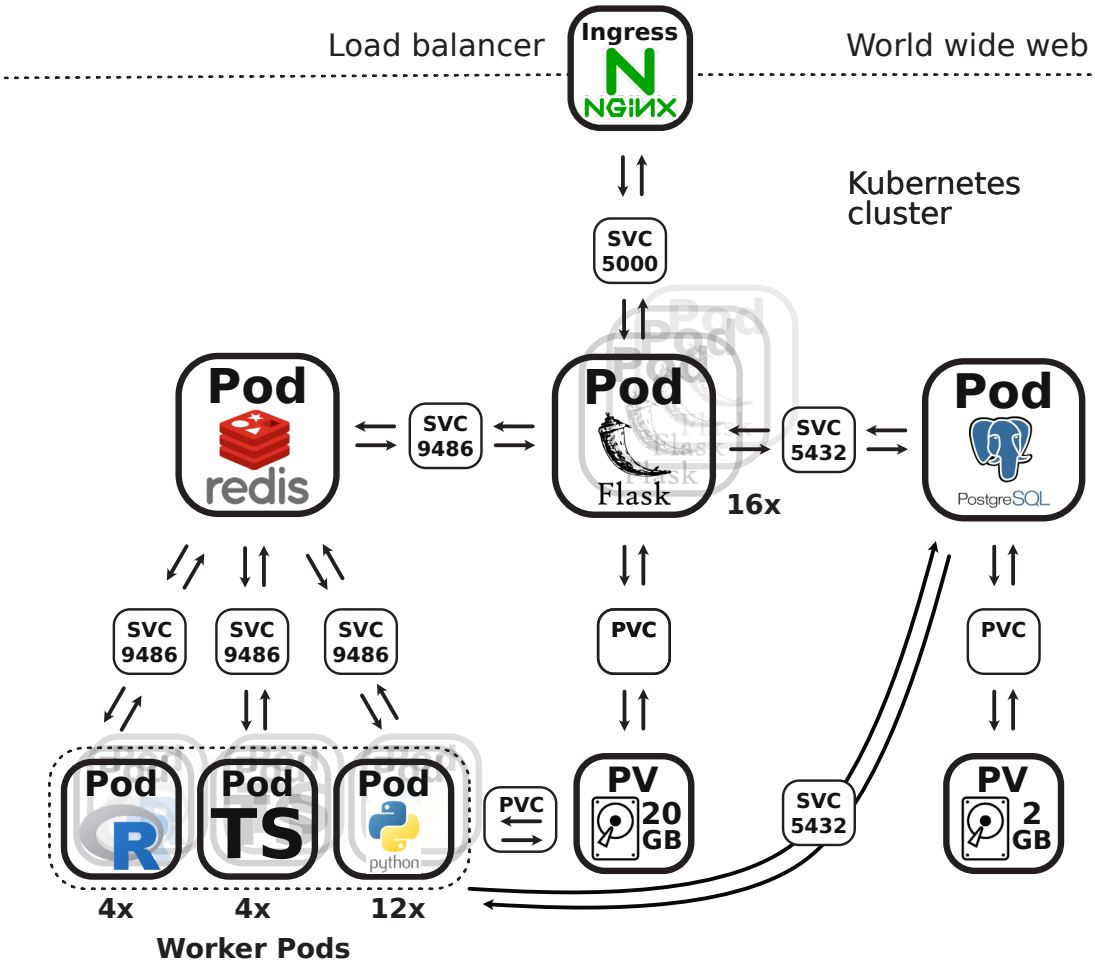


Fig. 2.16 **The TensorSignaturesOnline stack.** This schema represents the stack of the TensorSignatureOnline web app. Users accessing the application via the world wide web hit the NGINX load balancer which routes the user to one of the flask pods serving the front and backend of the website. To enable processing of many user requests at the same time, Flask pods are scaled horizontally with 16 replicas. Flask pods may access a PostgreSQL database and a Redis task queue via indicated services (SVC), as well as a persistent volume (PV) via the indicated persistent volume claim (PVC), which allows the Flask backend to store user uploaded VCF data. The PostgreSQL abstracts all data required to render the web site and stores these on a separate persistent volume. The Redis task queue coordinates the data processing pipeline which involves three different types of worker pods, each required to perform specific sub steps, i.e. the R Pod converts user uploaded VCF data to a data tensor, the TS Pod runs the tensor signature inference, and python pods plot the results of the analysis. Each of these "worker" pods are scaled horizontally to enable parallel processing of several user requests at a time and report back to the SQL database about the success or failure of a particular job.

collection of samples (route: `/create_dataset`); upload samples in VCF format to a

dataset (route: `/upload_sample/<dataset_id>`); or delete a dataset from the service (route: `/delete/<dataset_id>`). After creating and uploading samples to a dataset, users may start the analysis by submitting it to the data processing and analysis pipeline (route: `/analysis/<dataset_id>`) which triggers the submission of multiple jobs to a Redis task queue that ensures that the pipeline is executed step by step (Sec. 2.4.3). Moreover, I implemented a notification site which lists a detailed description of all submitted jobs, enabling users to report failures in case of their occurrence.

Fig. 2.17 **Views of the TensorSignatures web application.** **a**. The welcome page of TensorSignaturesOnline. **b**. The overview page with all extracted tensor signatures and a short description. **c**. Detailed signature page with SNV and other mutation type spectra as well as genomic properties. **d**. Cancer overview page with links to each sample of the respective cancer. **d**. Detailed sample page with true and predicted mutation type spectra, and corresponding exposures. **e**. Dashboard page which allows users to create datasets, upload samples in VCF format, and perform a TensorSignature analysis. This site becomes available after registering to the service.

A PostgreSQL database keeps track of users, datasets and samples

The SQL database can be considered as the heart of the application, as it represents the abstraction for users, datasets and samples, and tracks the progress of the data processing pipeline in respective tables. After registering to the application, a new user entry is created in the `users` table. This entry contains user specific information such as its `username` and `email` etc., and provides foreign keys to the tables `tasks`, `notifications` and `datasets`, which keep track of the user's uploaded data and submitted jobs (Fig. 2.18).

Upon login to the web application, the flask backend grants access to the dashboard route, which enables users to view their uploaded data and to create new datasets. The latter creates a new entry with the fields `name`, `path`, `processed` and `analysed` in the `datasets` table. The `name` field is a user specified name for the dataset, `path` stores path on the persistent volume to which all uploaded samples (VCFs) are saved to, and `processed` and `analysed` represent boolean variables that are set to `true` by the Redis task queue after the completion of certain steps in the data processing pipeline (Fig. 2.18). The dashboard displays datasets and gives access to an upload form which allows users to add samples in VCF format to an user owned dataset. The relationship between a `dataset` and a `sample` is represented via the `dataset` foreign key which is associated with each `sample` entry. After uploading one or many samples, a dataset may be subjected to the data processing pipeline, which converts the data to a count tensor and subsequently triggers the analysis with TensorSignatures.

A Redis task queue ensures proper processing of user data

The Redis pod coordinates the data processing pipeline which comprises three major steps: data preparation, inference and visualisation (Fig. 2.16). Each step of the pipeline runs within a dedicated microservice, because each subtask relies on a different set of dependencies and libraries. For example, the data preparation step is conducted within a pod that is running R, as it requires Bioconductor packages to convert the user uploaded VCFs into a mutation count tensor (2 GB); the inference is running in a Python container with Tensorflow (2 GB); while the data visualisation step is performed within a microservice equipped with plotting libraries (300-400 MB). Dedicating a single pod for each subtask enables to keep the size of each pod small, allowing for resource-saving horizontal scaling and efficient parallel processing of several jobs at a time.

Upon user submission of a dataset to the analysis pipeline, the Flask backend creates a series of `task` entries in the `tasks` table, often many for each subtask of the pipeline. For example, processing uploaded VCFs to a count tensor involves creating a variant tensor and computing trinucleotide frequencies; running TensorSignatures requires to run the inference

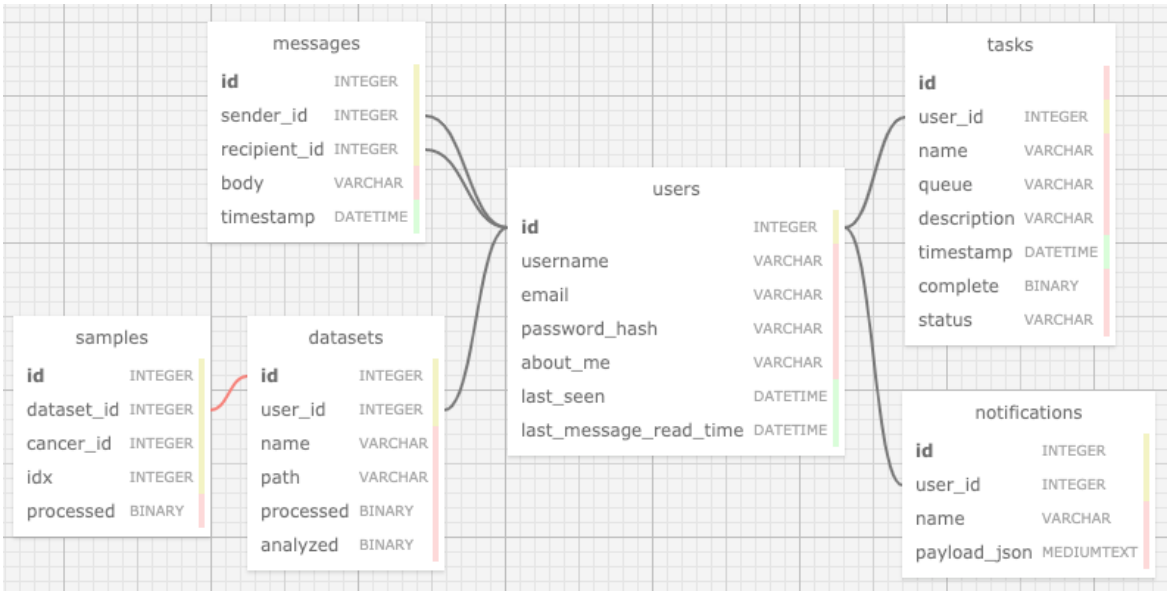


Fig. 2.18 **Database structure of the TensorSignaturesOnline web application.** The web application manages users and their data in four tables. The user table contains mainly user information and provides foreign keys to the the messages, tasks, datasets and notifications tables, thus linking each entry in the respective table uniquely to a user. The datasets table itself provides its id to each uploaded VCF file, thereby linking each sample each to a dataset.

and to normalise computed exposures; and finally plots need to be generated for each individual sample of the dataset. Like each entry of the `datasets` table, a `task` entry is uniquely associated to the user who submitted the request, and contains additional fields that specify the type of job, the queue on which the task should run, and a status and completion field. At the same time, each of these tasks are placed on the Redis task queuing system. As soon as an appropriate worker pod has freed up resources, it gets assigned with one of the outstanding tasks on the queue. Each of the worker pods run independently from the rest of the application, and report back to the queue about the completion or failure of the task. At this point the `complete` field of the respective `task` entry is set to true, and the `status` to ”success” or “failure”. Additionally, the status of each task is displayed in the notifications page of the dashboard, allowing users to keep track of the outstanding processes that have to be completed.

Also, to keep users informed about the progress of the analysis, a job notification is sent to their dashboard after the completion of each major step. These notifications are highlighted in the notification panel to enhance the user experience and provide a link to contact the Administrator in case of unexpected failures in the pipeline. This functionality is enabled by the `notification` and `message` tables of the database.

To ensure that each step runs in the right order, e.g. it does not make sense to run the inference without having the data tensor yet, jobs are scheduled with dependencies. As soon as a subprocess finishes, and the worker pod reports "success" or "failure" to the queue, the next job is started or all dependent jobs are cancelled, respectively. When all jobs associated with the submission of one dataset finish successfully, the frontend displays the results of the analysis for each sample, and allows the user to download the created data tensor in HDF5 format and the results from the TensorSignatures analysis.

Chapter 3

Results

While previous chapters mainly focused on methodological considerations and explained theoretical approaches to characterise mutational signatures beyond their 96 single base substitution spectra, the following focuses on applying the presented methodology to cancer genomes to learn the genomic properties of various mutational processes. To this end, I will present the analysis of 2,778 cancer genomes from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium (Sec. 3.1), as well as a validation analysis conducted on the Hartwig Medical Foundation (HMF) dataset (Sec. 3.2).

Contributions

This chapter is mainly based on the bioarxiv manuscript "Learning Mutational Signatures and their Genomic Properties with TensorSignatures" by Harald Vöhringer, Arne van Hoeck, Edwin Cuppen and Moritz Gerstung. H.V. conducted all bioinformatic analyses and produced the figures. A.v.H. and E.C. curated HMF data and provided computing resources for HMF data analysis by H.V.. M.G. conceived and supervised the analysis and developed code for categorising mutations. H.V. and M.G. wrote the manuscript with input from A.v.H. and E.C.

3.1 Discovering tensor signatures in the PCAWG dataset

We performed our discovery analysis on the PCAWG dataset, which is a international collaboration with the goal to identify common patterns of mutations in more than 2,600 cancer whole genomes from the International Cancer Genome Consortium. Section 3.1.1 briefly recapitulates the most important key points of the approach, followed by an overview of discovered tensor signatures in Sec. 3.1.2. In section 3.1.3, I will address two distinct mutational signatures of UV-light exposure found in active and quiescent chromatin, which

may be attributed to differential activity of nucleotide excision repair, and discuss in Sec. 3.1.4 transcription-associated mutagenesis manifesting as A[T>C] mutations. Sec. 3.1.5 discusses APOBEC mutagenesis, which manifests in two signatures reflecting highly clustered, double strand break repair initiated and lowly clustered replication-driven mutagenesis. The final section 3.1.6 of this chapter dives into somatic hypermutation, which produces a strongly clustered, TSS-associated signature in lymphoid cancers, and is distinct from a weakly clustered TLS signature found in multiple tumour types.

3.1.1 TensorSignatures jointly decomposes mutation spectra and genomic localisation

Here we analysed the somatic mutational catalogue of the PCAWG cohort comprising 2,778 curated whole-genomes from 37 different cancer types containing a total of 48,329,388 SNVs, 384,892 MNVs, 2,813,127 deletions, 1,157,263 insertions and 157,371 SVs.

Multiple mutation types contribute to mutagenesis

Data driven discovery of mutational signatures requires a meaningful classification of mutation types. We adopted the convention of classifying single base substitutions by expressing the mutated base pair in terms of its pyrimidine equivalent (C>A, C>G, C>T, T>A, T>C and T>G) plus the flanking 5' and 3' bases (Tab. B.1). We categorised other mutation types into 91 MNV classes (Tab. B.2), 62 insertion and deletion (indel) classes (Tab. B.3), and used the classification of SVs provided by the PCAWG Structural Variants Working Group (Tab. B.4, Li et al. (2020)).

Multidimensional genomic features produce a data tensor

Matrix-based mutational signature analysis proved to be powerful in deconvolving mutational spectra into mutational signatures, yet it is limited in characterising them with regard to their genomic properties. This is because individual mutations cannot always be unambiguously assigned *post hoc* to a given mutational process, which reduces the accuracy of measuring the genomic variation of closely related mutational processes. To overcome this limitation, we use 5 different genomic annotations – transcription and replication strand orientation, nucleosomal occupancy, consensus epigenetic state as well as local hypermutation – and generate 96-dimensional base substitution spectra for each possible combination of these genomic states separately and for each sample. Partitioning variants creates a seven-dimensional count tensor (a multidimensional array), owing to the multitude of possible combinations of different genomic features (Fig. 3.1).

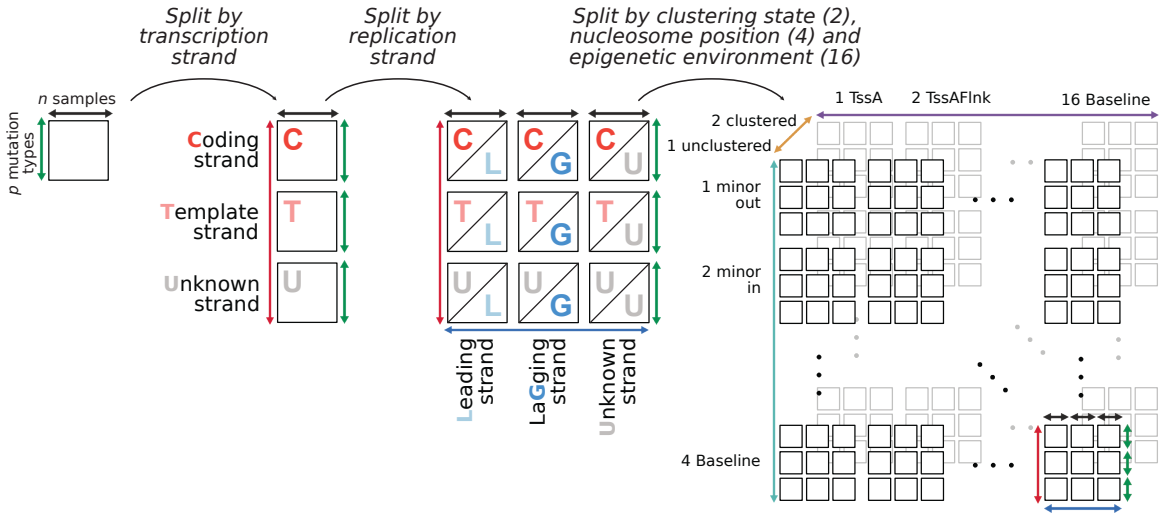


Fig. 3.1 **Splitting variants by transcriptional and replicational strand, and genomic states creates a multidimensional tensor.** Splitting variants by transcriptional and replicational strand, and genomic states creates an array of count matrices, a multidimensional tensor, in which each matrix harbours the mutation counts for each possible combination of genomic states.

Directional effects Mutation rates may differ between template and coding strand, because RNA polymerase II recruits transcription coupled nucleotide excision repair (TC-NER) upon lesion recognition on transcribed DNA only. Thus, TC-NER leads to lower mutation rates on the template strand, which is best illustrated by UV-induced mutations found in skin cancers (Pleasance et al., 2010a). TC-NER usually decreases the number of mutations in highly transcribed genes, but also the opposite effect – transcription coupled mutagenesis (TAM) – occurs (Haradhvala et al., 2016; Letouzé et al., 2017).

Similar to transcriptional strand asymmetries, mutation rates and spectra may differ between leading and lagging strand in replication (Haradhvala et al., 2016; Tomkova et al., 2018). This may be related to the fact that the leading strand is continuously synthesised by Pol ε, while lagging strand DNA synthesis is conducted by Pol δ, and is discontinuous due to formation of Okazaki fragments. Therefore, deficiencies in components involved in, or mutational processes interfering with DNA replication may lead to differential mutagenesis on leading or lagging strand.

Since not all mutations can be oriented either due to absent or bidirectional transcription, or because of unknown preferred replication direction far from a replication origin, this creates a total of 3×3 = (template, coding, unknown) $\times$ (leading, lagging, unknown) combinations of orientation states in the count tensor (Fig. 3.1).

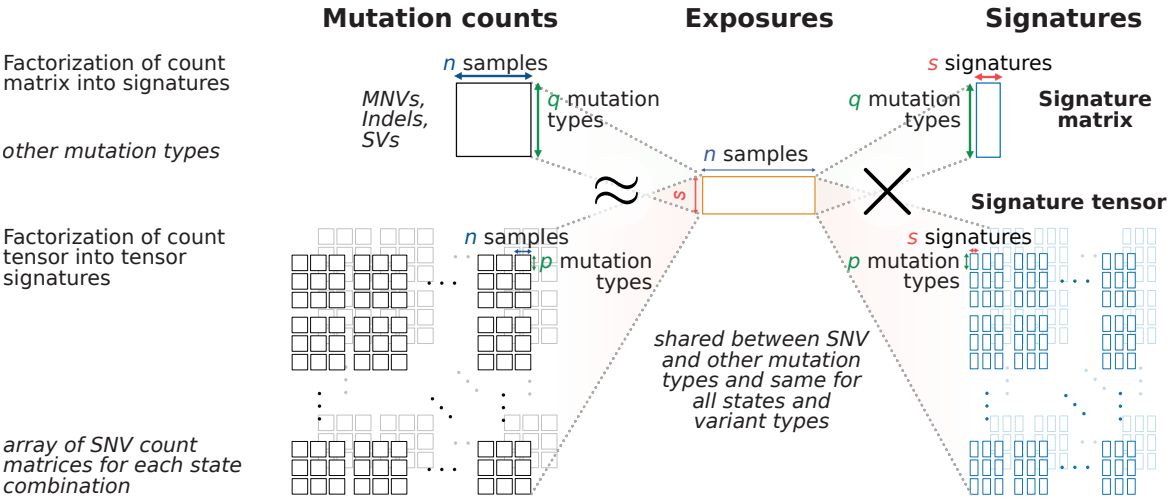


Fig. 3.2 **TensorSignatures factorises a mutation count tensor (SNVs) into an exposure matrix and signature tensor.** TensorSignatures factorises a mutation count tensor (SNVs) into an exposure matrix and signature tensor. Simultaneously, other mutation types (MNVs, indels, SVs), represented as a conventional count matrix are factorised using the same exposure matrix

(Epi-)genomic localisation factors Numerous studies found a strong influence of chromatin features on regional mutation rates. Strikingly, these effects range from the 10 bp periodicity on nucleosomes to the scale of kilo to mega bases caused by the epigenetic state of the genome (Pich et al., 2018; Schuster-Böckler and Lehner, 2012). To understand how mutational processes manifest on histone-bound DNA, we computed the number of variants on minor groove DNA facing away from and towards histone proteins, and linker DNA between two consecutive nucleosomes (Sec. 2.2.1). Additionally, we utilised ChromHMM annotations from 127 cell-lines to define epigenetic consensus regions (Sec. 2.2.1), which we used to assign SNVs to epigenetic contexts (Ernst and Kellis, 2012; Kundaje et al., 2015). Together this adds two dimensions of size 4 and 16 to the count tensor (Fig. 3.1).

Currently available data does not allow to obtain a comprehensive set of epigenetic annotations matched to the cell of origin for every cancer type. Using a consensus annotation assigns many regions to a variable (NA) state. While it is currently possible to match 31/37 PCAWG cancer types to cell lines closely corresponding to the presumed cell of origin, we note that 9,870,018 / 48,329,388 mutations change state using this partially matched annotation (Fig. B.2). We tested whether this has a major impact on the inference of epigenetic signature activities, which revealed that partially matched annotated epigenetic states do have good concordance with the all tissue consensus (Fig. B.3).

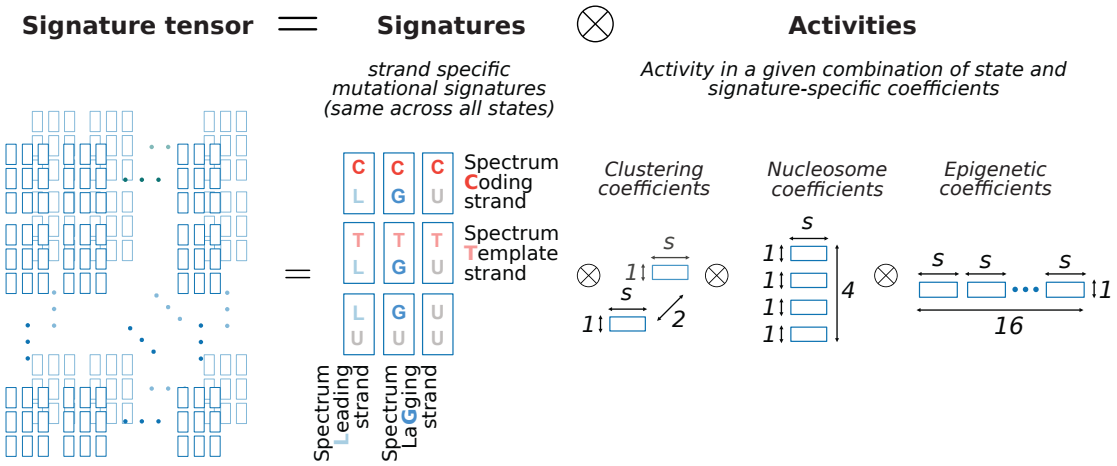


Fig. 3.3 **The lower dimensional structure of the signature tensor.** The signature tensor has itself a lower dimensional structure, defined by the product of strand-specific signatures, and coefficients reflecting the activity of the mutational process in a given genomic state combination.

Clustered mutations Finally, there are mutational processes capable of introducing large numbers of clustered mutations within confined genomic regions. This phenomenon is termed kataegis and is thought to be caused by multiple mutational processes (Nik-Zainal et al., 2012; Supek and Lehner, 2017). To detect such mutations, we developed a hidden markov model (HMM, Sec. 2.2.1) to assign the states clustered and unclustered to each mutation based on the inter-mutation distance between consecutive mutations. Separating clustered from unclustered mutations adds the final dimension in the mutation count tensor, which has a total of 6 dimensions with $2 \times 576 = 1{,}152$ combinations of states (Fig. 3.1).

TensorSignatures learns signatures based on mutation spectra and genomic properties

At its core, mutational signature analysis amounts to finding a finite set of prototypical mutation patterns and expressing each sample as a sum of these signatures with different weights reflecting the variable exposures in each sample. Mathematically, this process can be modelled by non-negative matrix factorisation into lower dimensional exposure and signature matrices. TensorSignatures generalises this framework by expressing the (expected value of the) count tensor as a product of an exposure matrix and a signature tensor (Fig. 3.2; Methods). The key innovation is that the signature tensor itself has a lower dimensional structure, reflecting the effects of different genomic features (Fig. 3.3). This enables the model to simultaneously learn mutational patterns and their genomic context – by drawing information from the whole dataset even when the number of combinations of genomic states becomes high (1,152), thus yielding a more accurate inference to conventional NMF

relying on a 96-trinucleotide channel decomposition only and subsequent assessment of signature properties (Sec. 2.3.2, Fig. 2.11). In this parametrisation each signature is represented as a set of 2×2 strand-specific mutation spectra and a set of defined coefficients, measuring its activity in a given genomic state of a given dimension. TensorSignatures incorporates the effect of other variants (MNVs, indels, SVs), which remain unoriented and are expressed as a conventional count matrix, by sharing the same exposure matrix as SNVs, thus enabling to jointly learn mutational processes across different variant classes more robustly in comparison to approaches which rely on (post-hoc) matching mutational spectra (Sec. 2.3, Fig. 2.12). TensorSignatures models mutation counts with an over-dispersed negative binomial distribution, which we tested extensively on simulated data sets (Sec. 2.3.1, Fig. 2.8-2.10), and enables to choose the number of signatures with established statistical model selection criteria, such as the Bayesian Information Criterion (BIC, Sec. 2.2.5, Fig. 2.7, B.1).

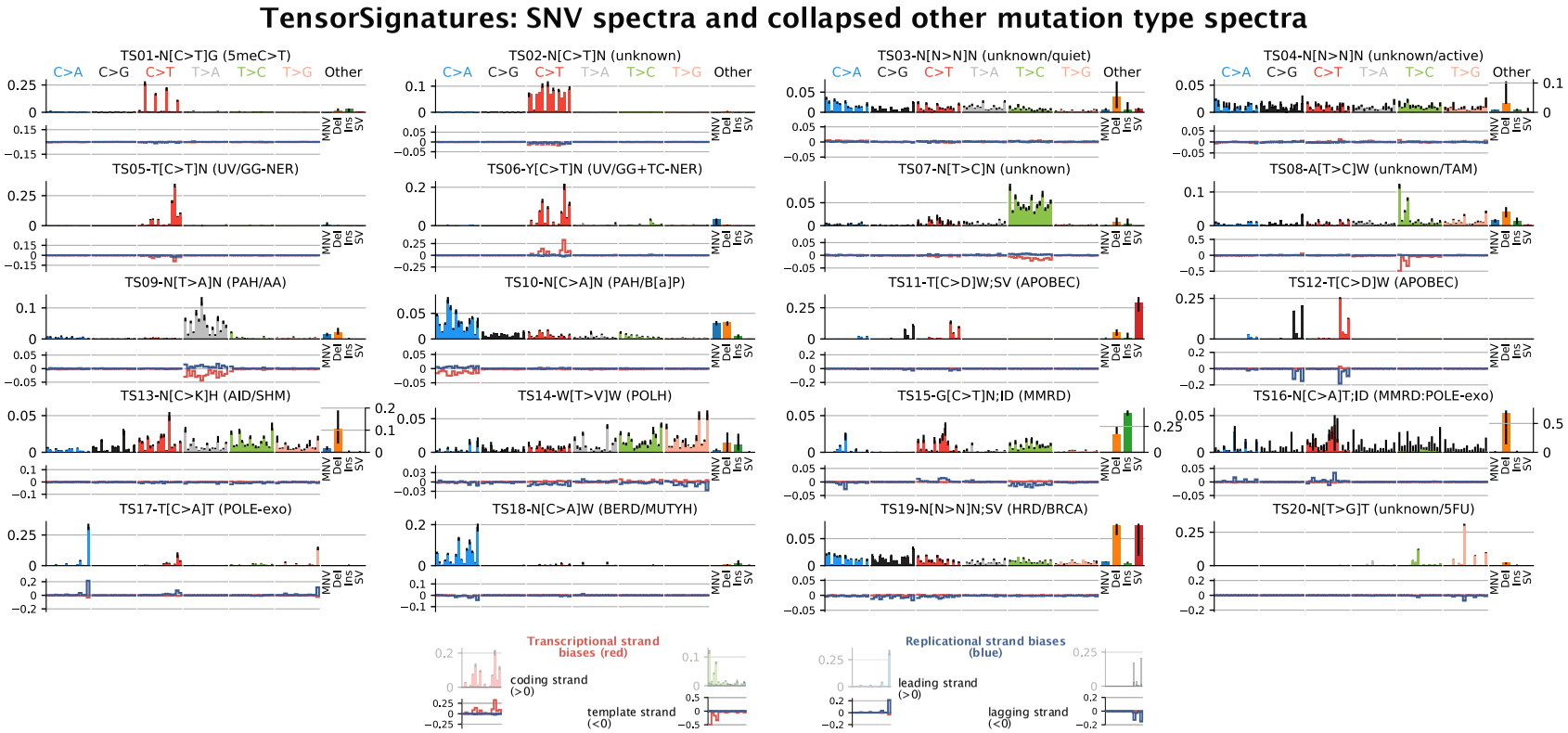


Fig. 3.4 **The analysis of 2778 whole genomes revealed 20 tensor signatures.**. Upper panels depict SNV spectra, and a summarized representation of associated other mutation types. SNV mutations are shown according to the conventional 96 single base substitution classification based on mutation type in a pyrimidine context (color) and 5' and 3' flanking bases (in alphabetical order). The panel under each SNV spectrum indicates transcriptional (red), and replicational strand biases (blue) for each mutation type, in which negative deviations indicate a higher probability for template or lagging strand pyrimidine mutations, and positive amplitudes a larger likelihood for coding or lagging strand pyrimidine mutations (and vice versa for purine mutations).

3.1.2 Mutational Signatures are composed of a multitude of mutation types and vary across the genome

Analysis of 2778 genomes produces 20 TensorSignatures

Applying TensorSignatures to the PCAWG dataset and using the conservative BIC (Fig. S2) produced 20 tensor signatures (TS) encompassing mutational spectra for SNVs and other mutation types (Fig. 3.4), and associated genomic properties (Fig. 3.5). Reassuringly, we extracted a number of signatures with SNV spectra highly similar to the well curated catalogue of COSMIC signatures (Tab. B.5, Alexandrov et al. (2018); Forbes et al. (2015)). Interestingly, our analysis revealed a series of signatures that have similar SNV spectra in common, but differ with regard to their genomic properties or mutational composition. These signature splits indicate how mutational processes change across the genome and will be discussed in further detail below. In the following, we refer to signatures via their predominant mutation pattern and associated genomic properties. Of the 20 signatures, 4 were observed in nearly every cancer type: TS01-N[C>T]G, characterised by C>T mutations in a CpG context, most likely due to spontaneous deamination of 5-meC, similar to COSMIC SBS1, TS02-N[C>T]N of unknown aetiology, and two signatures with relatively uniform base substitution spectra, TS03-N[N>N]N (unknown/quiet chromatin), and TS04-N[N>N]N (unknown/active chromatin), which loosely correspond to SBS40 and SBS5 (Fig. 3.6).

Signatures are defined by diverse mutation types and genomic properties

While the most prevalent mutations are single base substitutions, there are 16/20 signatures with measurable contributions from other mutation types (>1 %; Fig. 3.5). The most notable cases are TS15-G[C>T]N;ID, which is similar to a compound of COSMIC signatures SBS6/15/26 + ID1/2 and characterised by C>T transversions in a GCN context and frequent mono-nucleotide repeat indels (Fig. C.73) indicative of MMRD. Similarly, TS16-N[C>A]T;ID, likely to reflect concurrent MMRD and POLE exonuclease deficiency, exhibits large probabilities for deletions (Fig. C.78) and a base substitution pattern similar to SBS14. Large proportions of SVs ($\sim$25 %) were found in TS11-T[C>D]W;SV (D = A, G, or T; W = A or T), which reflects SV-associated APOBEC mutagenesis caused by double strand break repair with a base substitution spectrum similar to SBS2/13. Furthermore, TS19-N[N>N];SV apparently reflects a pattern of homologous recombination deficiency (HRD), characterised by a relatively uniform base substitution pattern similar to SBS3, but a high frequency of SVs, in particular tandem duplications (Fig. C.93).

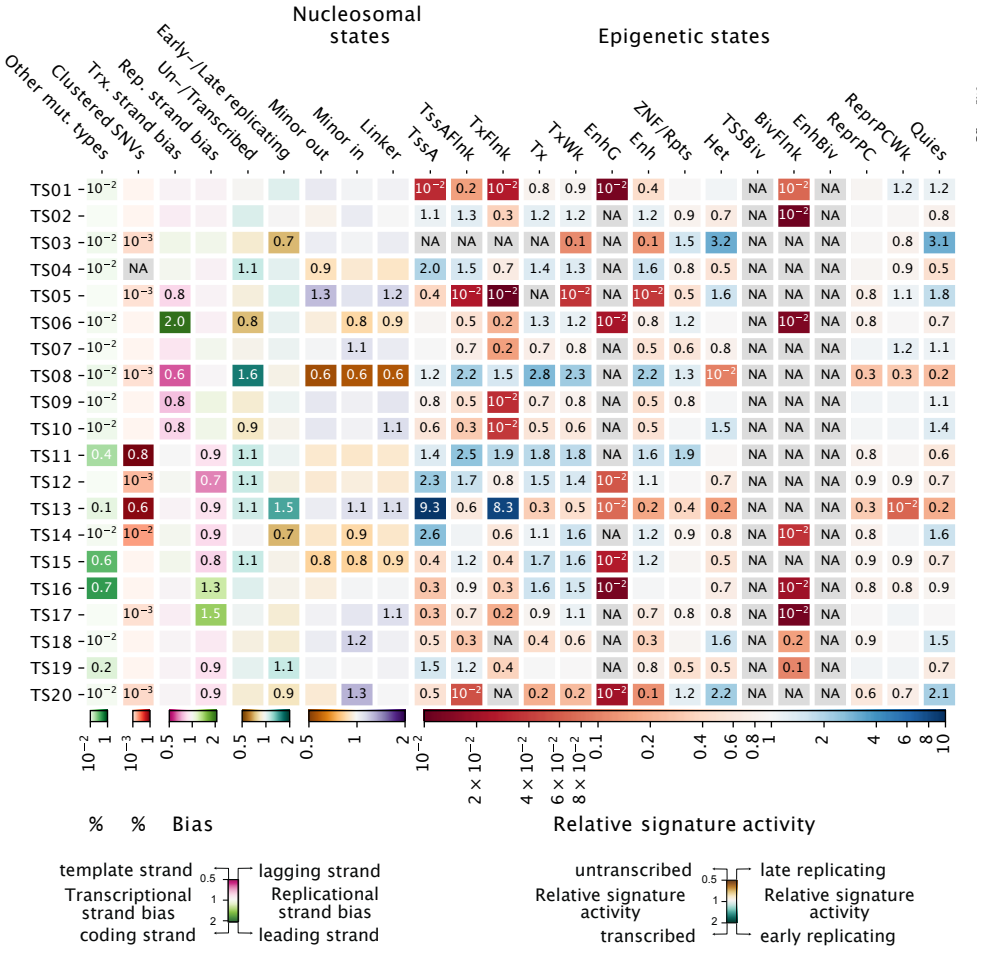


Fig. 3.5 **Tensor factors describe a multitude of genomic properties of each tensor signature.** Heatmap visualisation of extracted tensor factors describing the genomic properties of each tensor signature. Proportions of other mutation types and clustered SNVs are indicated in percentages. Transcriptional and replicational strand biases indicate shifts in the distribution of pyrimidine mutations on coding/template and leading/lagging strand. Coefficients < 1 (pink) indicate signature enrichment on template or lagging strand DNA, and conversely values > 1 (green), a larger mutational burden on coding or leading strand (a value of 1 indicates no transcriptional or replicational bias). Relative signature activities in transcribed/untranscribed and early/late replicating regions. Coefficients > 1 (turquoise) indicate enrichment in transcribed and early replicating regions, while values < 1 (brown) indicate a stronger activity of the mutational process in untranscribed or late replicating regions. Relative signature activities on nucleosomes and linker regions, and across epigenetic states as defined by consensus ChromHMM states. Scores indicate relative signature activity in comparison to genomic baseline activity. A value of 1 means no increase or decrease of a signature's activity in the particular genomic state, while values > 1 indicate a higher, and values < 1 imply a decreased activity.

APOBEC associated signatures TS11 and TS12 as well as mutational processes TS13 and TS14 due to AID activity induce clustered mutations

9/20 signatures displayed a measurable propensity to generate clustered mutations (>0.1 %; Fig. 3.5). The proportions of clustered mutations produced by each mutational process were highest in signatures associated with APOBEC and activation-induced deaminase (AID) activity: Up to 79 % and 0.6 % of SNVs attributed to TS11-T[C>D]W;SV and TS12-T[C>D]W, respectively, were clustered, with otherwise indistinguishable base substitution spectra. A similar phenomenon was observed in two signatures reflecting Pol η driven SHM. While both TS13-N[C>K]H (K = G or T; H = A, C, or T) and TS14-W[T>V]W (V = A, C, or G) have only mildly diverging base substitution spectra, with TS14 being similar to SBS9, they dramatically differ in the rates at which they generate clustered mutations, which are 59 % and 1 %, respectively (Fig. 3.5).

UV-light associated TS06 and an unknown mutational process TS08 linked to A[T>C]W mutagenesis predominantly affect coding and template strand DNA

5/20 signatures exhibit substantial transcriptional strand bias (TSB $\geq$10 %; Fig. 3.5). This is strongest in the UV-light associated signature TS06-Y[C>T]N (Y = C or T), similar to SBS7b, where the rate of C>T substitutions on the template strand was half of the corresponding value on the coding strand, highly indicative for active TC-NER. In contrast, TS08-A[T>C]W, similar to SBS16, shows largest activities in liver cancers and preferably produces T>C transitions on template strand DNA. In line with a transcription-coupled role, the activity of TS08 shows a noteworthy elevation in transcribed regions. Both signatures will be discussed in more detail later on (Sec. 3.1.3 and 3.1.4).

Mutational processes TS12 and TS17 linked to APOBEC and POLE activity introduce single base substitution with strong replicational strand biases

Analysis of pyrimidine/purine shifts in relation to the direction of replication indicated 9/20 signatures with replication strand biases (RSB $\geq$10%). In accordance with previous studies, TS12-T[C>D]W asserts a higher prevalence of APOBEC-associated C>D mutations, consistent with cytosine deamination, on lagging strand DNA which is thought to be exposed for longer periods as opposed to more processively synthesised leading strand DNA. Conversely, TS17-T[C>A]T, associated with POLE exonuclease variants (SBS10a/b), displays a pyrimidine bias towards the leading strand (Fig. 3.5, Haradhvala et al. (2018)). Since Pol ε performs leading strand synthesis, the strand bias indicates that C>A (G>T) mutations arise on a template C, presumably through C·dT misincorporation (Shinbrot et al., 2014). Further

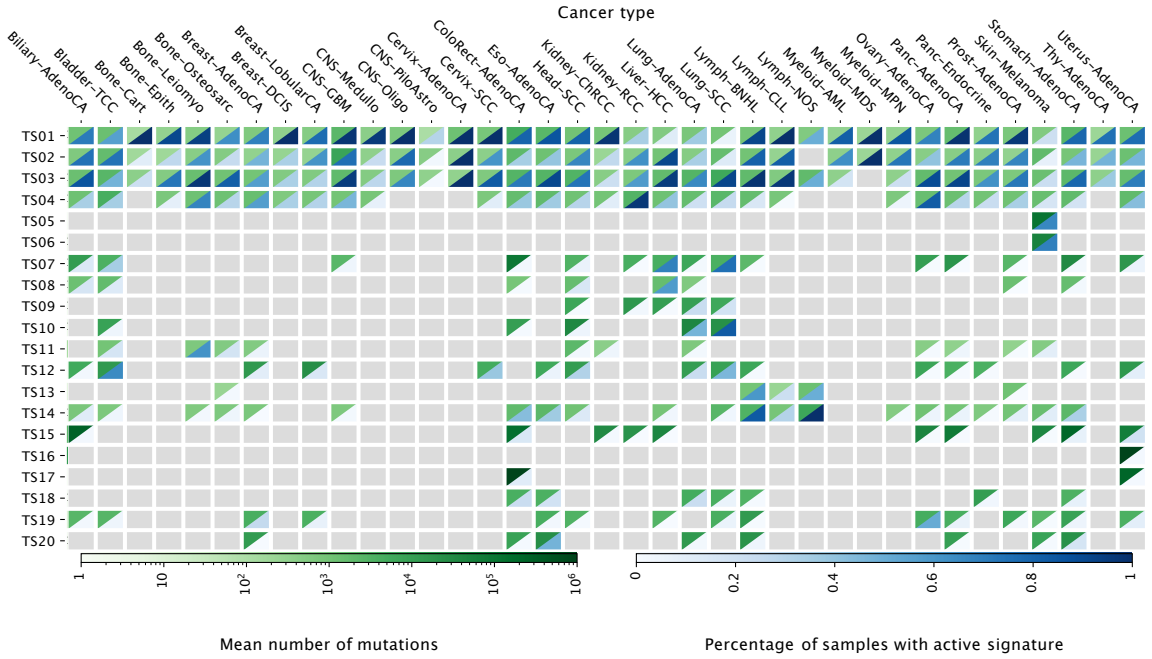


Fig. 3.6 **Signature activity in different cancer types (Exposures).** Signature activity in different cancer types (Exposures). Upper triangles (green) indicate the mean number of mutations contributed by each signature, lower triangles show the percentage of samples with a detectable signal of signature defined as the number of mutations attributed to the signature falling into a signature-specific typical range (Methods). Greyed boxes indicate cancer types for which a signature was not found to contribute meaningfully.

examples with replication strand biases include the MMRD-associated signatures TS15 and TS16 discussed above. Of note, the two SHM-associated signatures TS13 and TS14 displayed opposing patterns with respect to their activity in oriented (early) and unoriented (late) replicating regions (Fig. 3.5).

Genomic localisation factors modulate signatures activities, with epigenetic states having the greatest influence

To understand how mutational processes manifest on nucleosomal DNA, we estimated signature activities on minor groove DNA facing away from and towards histone proteins, and linker DNA between two consecutive nucleosomes (Fig. 3.5). Almost all signatures showed either an increase or a decrease of mutational rates across all nucleosomal states. The only exception to this rule is TS20-N[T>G]T (SBS17a/b), which showed a slight decrease in the outward facing minor groove, while the inwards facing showed elevated mutation rates

(Pich et al., 2018). TS20 is likely caused by incorporation of dUTP or oxo-dTTP (Tomkova et al., 2018), possibly, but not necessarily, due to 5-FU treatment (Christensen et al., 2019).

Considering the activities of mutational processes across epigenetic domains, our analysis indicates that there is not a single mutational process which is acting uniformly on the genome (Fig. 3.5). However, our results suggest that mutational processes may be categorised into two broad groups: Those that are elevated in active (TssA, TssAFlnk, TxFlnk, Tx and TxWk) and depleted in quiescent regions (Het, Quies), and vice versa. This phenomenon includes the two omnipresent signatures with relatively uniform spectra TS03-N[N>N]N and TS04-N[N>N]N, suggesting a mechanism associated with the chromatin state behind their differential manifestation (Fig. 3.4). This also applies to two signatures associated with UV exposure, TS05-T[C>T]N and TS06-Y[C>T]N, and also two signatures of unknown aetiology, most prominently found in liver cancers, TS07-N[T>C]N, similar to SBS12, and TS08-A[T>C]W, which we will discuss in detail in the following section.

3.1.3 The spectrum of UV mutagenesis changes from closed to open chromatin

Two signatures, TS05-T[C>T]N and TS06-Y[C>T]N, were exclusively occurring in Skin-Melanoma and displayed almost perfect correlation (Spearman R^2 =0.98, Fig. B.4) of attributed mutations, strongly suggesting UV mutagenesis as their common cause. Both signatures share a very similar SNV spectrum, only differing in the relative extent of C[C>T]N and T[C>T]N mutations, which is more balanced in TS06 (Fig. 3.7). However, they strongly diverge in their activities for epigenetic contexts and transcriptional strand biases: TS05 is enriched in quiescent regions, and shows no transcriptional strand bias, while the opposite is true for TS06, which is mostly operating in active chromatin (Fig. 3.5). Of note, the spectra of these signatures closely resemble that of COSMIC SBS7a and SBS7b, which have been suggested to be linked to different classes of UV damage (Hayward et al., 2017). However, as our genomically informed TensorSignature inference and further analysis show, the cause for the signature divergence may be found in the epigenetic context, which seemingly not only determines mutation rates, but also the resulting mutational spectra.

Diverging transcriptional strand biases and base substitution spectra differentiate TS05 and TS06

A characteristic difference between the two signatures is the presence of a strong transcriptional strand bias in signature TS06, which is almost entirely absent in signature TS05 (Fig. 3.7). To verify that this signature inference is correct, and the observed bias and spectra

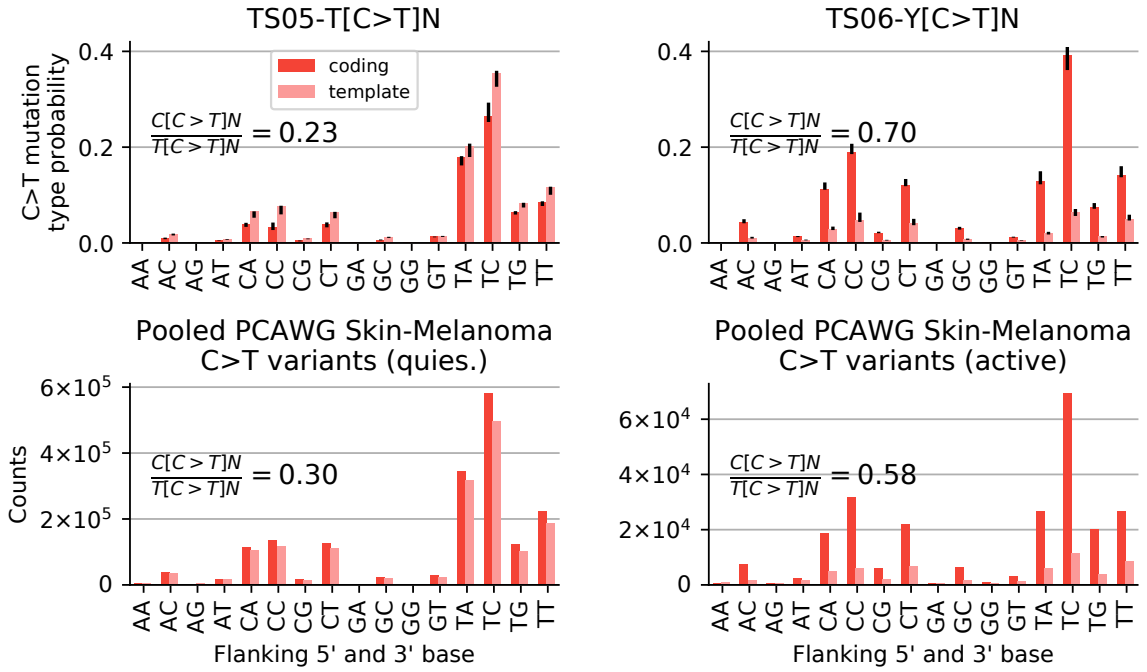


Fig. 3.7 **TS05 and TS06 spectra for coding and template strand DNA, and pooled PCAWG Skin-Melanoma C>T variant counts.** C>T mutation probabilities of TensorSignatures TS05 and TS06 for coding and template strand DNA (upper panel), and pooled PCAWG Skin-Melanoma C>T variant counts (lower panel) from coding and template strand DNA in epigenetically active (TssA, TssAFlnk, TxFlnk, Tx and TxWk, right) and quiescent regions (Het and Quies, left).

are genuinely reflecting the differences between active and quiescent chromatin, we pooled C>T variants from Skin-Melanoma samples which revealed that the data closely resembled predicted spectra (Fig. 3.7). In addition, quiescent chromatin also displays a predominant T[C>T]N substitution spectrum (5'C/5'T = 0.3), while the spectrum in active chromatin is closer to Y[C>T]N (5'C/5'T = 0.58), as predicted by the signature inference (Fig. 3.7). This difference does not appear to be related to the genomic composition, and holds true even when adjusting for the heptanucleotide context (Fig. B.5).

The spatial distribution of SNVs in Skin-Melanoma corroborate properties of TS05 and TS06

To verify this, we sought to understand the spatial distribution of mutations in Skin-Melanoma across the genome. Concretely, if properties of TS05 and TS06 prove well-founded, we would expect changes in the number of C>T mutations on coding and template strand DNA, as well as an increase of C[C>T]N mutations (spectral shift) as we move from quiescent to

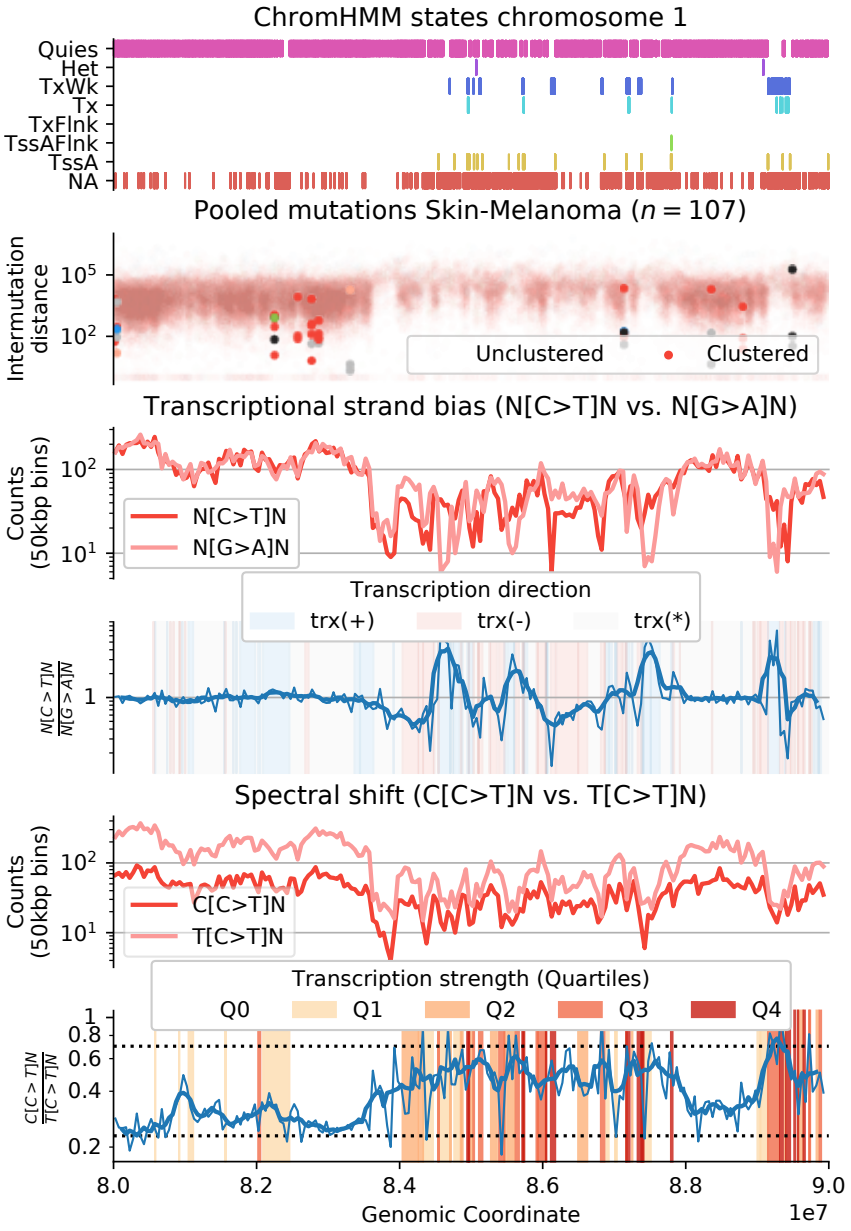


Fig. 3.8 **A spatial analysis of UV-mutagenesis in Skin Melanoma (*n*=107).** Upper panel: Consensus ChromHMM states from a representative 10 Mbp region on chromosome 1, and the corresponding mutational density of pooled Skin-Melanoma samples. Middle panel: N[C>T]N and N[G>A]N counts in 50kbp bins, and their respective ratios (thin blue line: ratio; thick blue line: rolling average over 5 consecutive bins) illustrate the transcriptional strand bias of C>T mutations in quiescent and active regions of the genome. Lower panel: Relationship between expression strength and the spectral shift of C>T mutations in terms of binned C>T variant counts in TpC and CpC context and their respective ratios (thin blue line) as well as a rolling average (thick blue line).

active genomic regions. To demonstrate this, we selected a representative 10 Mb region from chromosome 1 comprising a quiescent and active genomic region as judged by consensus ChromHMM states, and the varying mutational density from pooled Skin-Melanoma samples (Fig. 3.8, upper panel).

To visualise the transcriptional strand bias of UV mutagenesis, we counted the number of N[C>T]N and its purine equivalent N[G>A]N in 50 kb bins, because dependent on transcription directionality, these mutations reflect C>T substitutions on template or coding strand. Strikingly, our analysis revealed that numbers of N[C>T]N and N[G>A]N counts are roughly equal in quiescent regions, but start to fluctuate in active genomic compartments which is best visualised by their ratios (Fig. 3.8, middle panel).

Next, to test whether the distribution of C>T mutations changes from quiescent to active genomic regions, we counted pooled C[C>T]N and T[C>T]N variants in 50 kb bins (Fig. 3.8). While the difference of T[C>T]N and C[C>T]N mutations is largest in quiescent regions, it almost diminishes in transcribed genomic regions. We then hypothesised that the degree of expression may modulate the spectrum of C>T mutations, and computed therefore median gene expression quartiles from Skin-Melanoma samples ($n = 11$), which we show together with the ratio of C[C>T]N and T[C>T]N variants. Strikingly, this analysis showed that the ratio equals roughly 0.2 in quiescent regions but rises to approximately 0.7 in highly transcribed regions which closely resembles the prediction of TS05 and TS06 respectively (Fig. 3.8, lower panel).

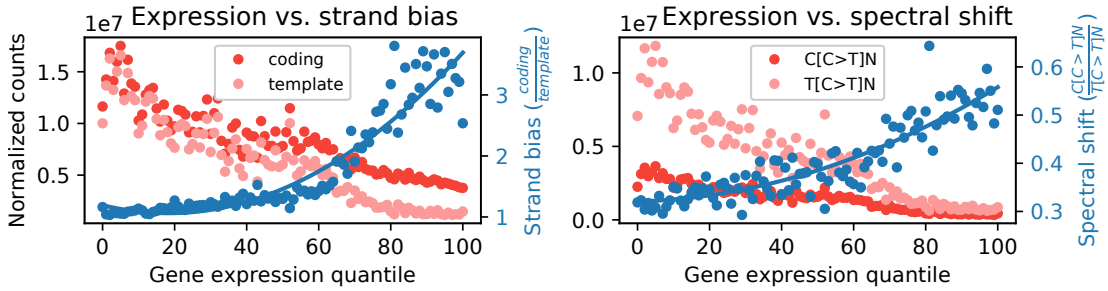


Fig. 3.9 **The effects of gene expression on the transcriptional strand bias and the mutational spectrum of C>T mutations in Skin-Melanoma (PCAWG, n=107).** Gene expression strength vs. transcriptional strand bias (measured by the ratio normalised C>T variants in Skin-Melanoma on coding and template strand), and gene expression strength vs. C[C>T]/T[C>T] spectral shift (indicated as the ratio of normalize C>T mutations in 5'C and 5'T context).

RNA-seq expression data substantiate the evidence for epigenetic modulation of UV-mutagenesis

These observations are further corroborated by RNA-seq data available for a subset of samples ($n = 11$): The transcriptional strand bias is most pronounced in expression percentiles greater than 50 leading to an increased ratio of coding to template strand mutations (Fig. 3.9). Again, the decline is accompanied by a shift in the mutation spectrum: While both C[C>T]N and T[C>T]N variant counts decline steadily as gene expression increases, the reduction of C[C>T]N mutations is larger in comparison to T[C>T]N mutations, which manifests as an increasing C[C>T]N and T[C>T]N ratio, reaching a ratio of approximately 0.5 in the highest expression quantiles (Fig. 3.9).

TS05 and TS06 may represent the interplay of UV-mutagenesis and different modes of nucelotide excision repair

The diverging activity in relation to the chromatin state suggests an underlying differential repair activity. Global genome nucleotide excision repair (GG-NER) clears the vast majority of UV-lesions in quiescent and active regions of the genome and is triggered by different damage-sensing proteins. Conversely, TC-NER is activated by template strand DNA lesions of actively transcribed genes. As TS05 is found in quiescent parts of the genome, it appears likely that it reflects the mutation spectrum of UV damage as repaired by GG-NER (Sec. 1.4.2). Based on the activity of TS06 in actively transcribed regions and its transcriptional strand bias, it seemingly reflects the effects of a combination of GG- and TC-NER, which are both operating in active chromatin. This joint activity also explains the fact that the spectrum of TS06 is found on both template and coding strands.

GG-NER deficient XPC$^{-/-}$ cutaneous squamous cell carcinomas genomes lack TS05

This attribution is further supported by data from $n = 13$ cutaneous squamous cell carcinomas (cSCCs) of $n = 5$ patients with Xeroderma Pigmentosum, group C, who are deficient of GG-NER and $n = 8$ sporadic cases which are GG-NER proficient (Zheng et al., 2014). XPC/GG-NER deficiency leads to an absence of TS05 in quiescent chromatin and to a mutation spectrum that is nearly identical in active and quiescent regions of the genome (Fig. 3.10). Furthermore, the UV mutation spectrum of XPC/GG-NER deficiency, which is thought to be compensated by TC-NER, differs from that of TS06, reinforcing the notion that TS06 is a joint product of GG- and TC-NER. This is further supported by the observation that XPC/GG-NER deficiency leads to a near constant coding strand mutation rate, independent of transcription strength (Fig. 3.10, Zheng et al. (2014)), indicating that the transcriptional

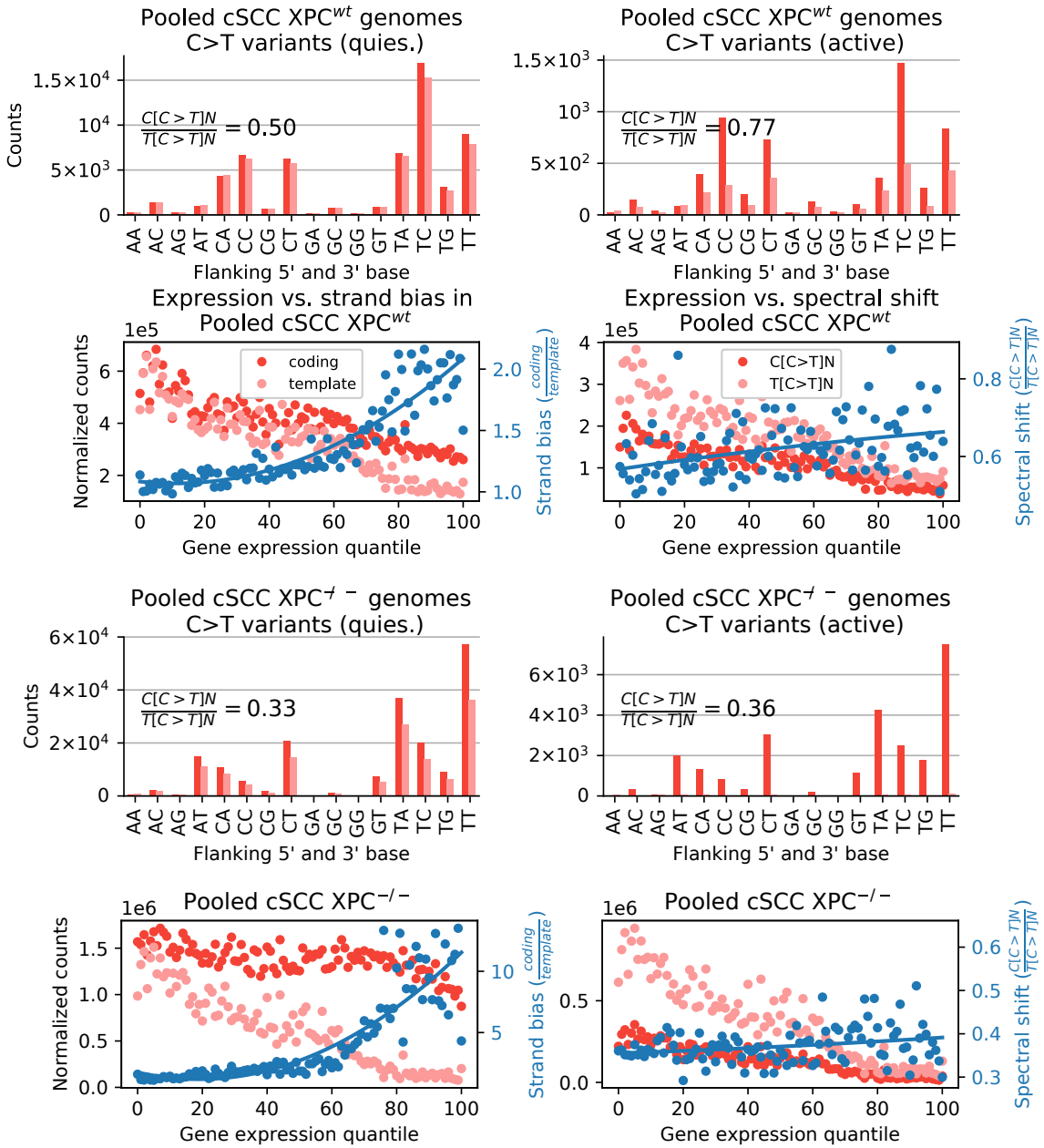


Fig. 3.10 **The spectrum C>T mutations of pooled XPC^{wt} and XPC$^{-/-}$ cSCC genomes.** Pooled cSCC C>T variant counts from coding and template strand DNA in epigenetically active (TssA, TssAFlnk, TxFlnk, Tx and TxWk, right) and quiescent regions (Het and Quies, left), as well as the relationship between gene expression and transcriptional strand bias (measured by the ratio normalised C>T variants), and the relation between gene expression and the C[C>T]/T[C>T] spectral shift (indicated as the ratio of normalised C>T mutations in 5'C and 5'T context) in GG-NER wildtype (XPC^{wt}, upper panel) and deficient (XPC$^{-/-}$, lower panel) cSCC genomes. Blue curves: quadratic fit.

dependence of coding strand mutations in GG-NER proficient melanomas and cSCCs is due to transcriptionally facilitated GG-NER.

While the activity patterns of TS05/06 and appear to be well aligned with GG-NER and GG/TC-NER, these observations, however, do not explain the observed differences in mutation spectra. The fact that the rates of C[C>T]N and T[C>T]N mutations change between active and quiescent chromatin – and the fact the these differences vanish under XPC/GG-NER deficiency – suggests that DNA damage recognition of CC and TC cyclobutane pyrimidine dimers by GG-NER differs between active and quiescent chromatin, with relatively lower efficiency of TC repair in quiescent genomic regions, as evidenced by TS05.

3.1.4 Transcription-associated mutagenesis manifests in an ApT context in highly transcribed genes

Diverging mutational spectra between active and quiescent chromatin were also observed in liver cancers (Fig. 3.5, 3.6), driven by differential activity of TS07-N[T>C]N and TS08-A[T>C]W, which closely resemble COSMIC signatures SBS12 and SBS16, respectively. In line with previous findings, there was a strong transcriptional bias of TS08, introducing 1.6× more T>C variants on the template strand (Fig. 3.5). While both signatures are most frequently found in liver cancers, where they are strongly correlated ($R^2 = 0.68$, Fig. B.6), they are also observed in a range of other cancers, indicating that they are reflecting endogenous mutagenic processes.

TS08 is characterised by a depletion of single base substitutions in 5'-B context and a strong transcriptional strand bias towards template strand DNA

The most prominent difference between these signatures is the depletion of mutation types in 5'-B context on coding strand DNA in TS08 (Fig. 3.11; B = C, G, or T). This attribution into signatures is confirmed when directly assessing mutation spectra in active and quiescent regions of Liver-HCC (Fig. 3.11). TS08 displays a strong transcriptional strand bias, as previously noted for SBS16 (Letouzé et al., 2017), and is confirmed here by a direct investigation of variant counts. A further defining feature of TS08 are indels $\geq$2 bp (Fig. 3.4, C.38), which were reported to frequently occur in highly expressed lineage-specific genes in cancer (Imielinski et al., 2017), consistent with experimental data of transcription-replication collisions (Sankar et al., 2016).

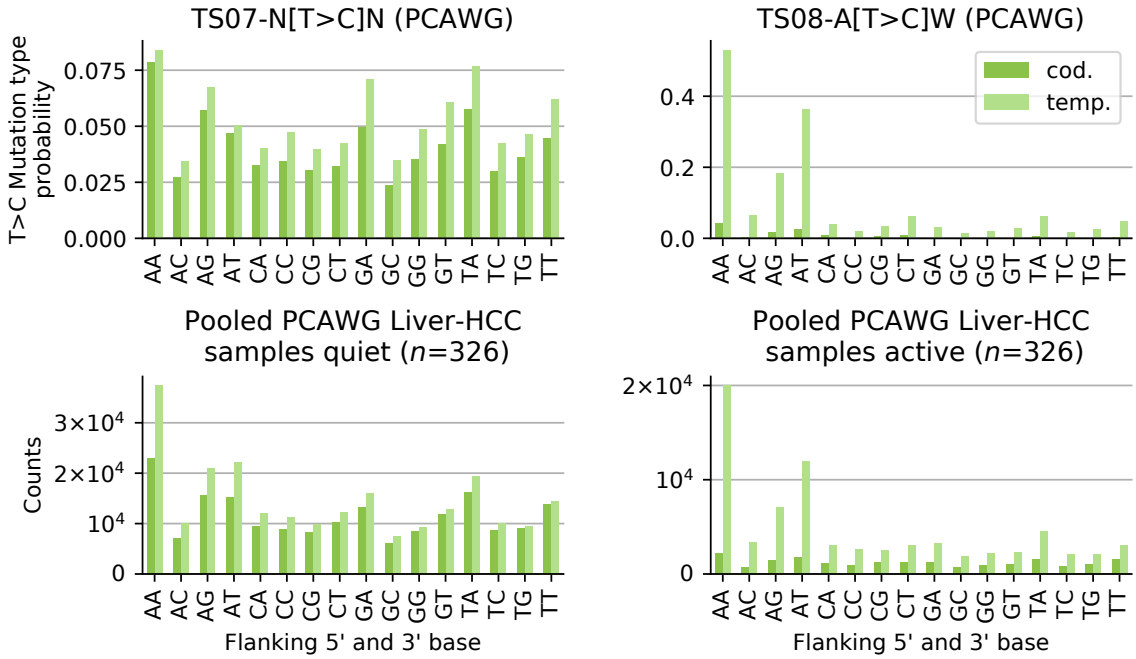


Fig. 3.11 **Spectral differences of T>C mutagenesis in liver cancers.** Upper panel: T>C mutation type probabilities of TensorSignatures TS07 and TS08 for coding and template strand DNA. Lower panel: Pooled PCAWG Liver-HCC T>C variant counts for coding and template strand DNA in epigenetically active and quiescent regions.

Active and quiescent genomic regions determine the interplay of the mutational processes TS07 and TS08 in Liver-HCC

In Liver-HCC, these two processes produce a regionally changing mutation spectrum between active and quiescent genomic environments (Fig. 3.12). Indeed, the ratio of T>C and complementary A>G mutations confirmed that the transcriptional strand bias of TS08 arises exclusively in active genomic regions (Fig. 3.12). These are accompanied by a change from a N[T>C]N and to an A[T>C]W spectrum, changing from a 5'A/5'B ratio of approximately 0.4 in quiescent regions to a value of up to 1 in active regions (Fig. 3.12, B.7).

Transcription-associated mutagenesis manifesting as A[T>C] mutations is found in a range of cancer types

We then systematically analysed the effect of transcription strength on T>C mutagenesis in samples with detectable TS07 and TS08 exposures in Liver-HCC and different cancer types. Mutation rates showed a dynamic relation to transcriptional strength (Fig. 3.13). Initially, normalised counts of T>C mutations on coding and template strand initially decline for low transcription. Yet this trend only continues on the coding strand for transcription

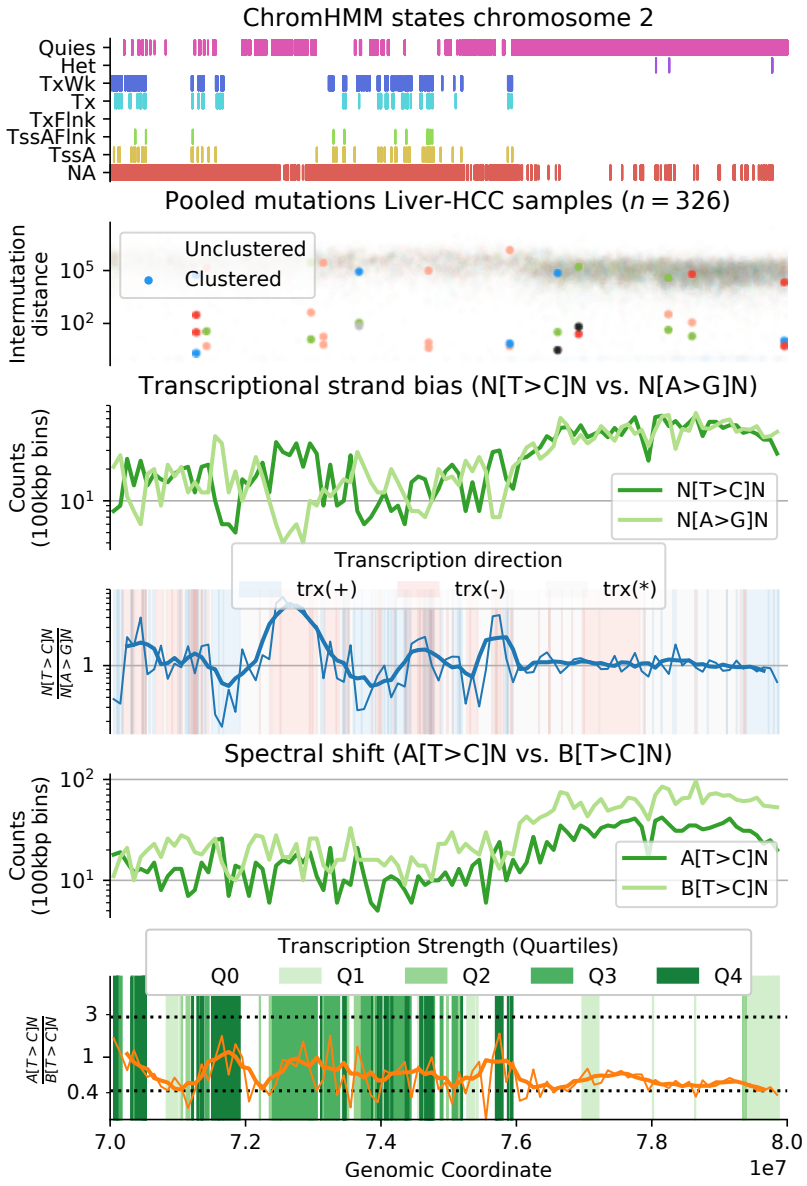


Fig. 3.12 **Spatial analysis of T>C mutagenesis in liver cancers.** Upper panel: Consensus ChromHMM states from a representative 10Mbp region on chromosome 2 depicting an active and quiescent genomic region, and the corresponding mutational density from pooled Liver-HCC samples. Middle panel: Illustration of the transcriptional strand bias in terms of 100kbp binned N[T>C]N and N[A>G]N counts, and respective ratio (thin blue line). The thick blue line depicts the corresponding rolling average over 5 consecutive bins. Lower panel: Changes in the distribution of T>C mutations in an active and quiescent genomic regions in terms of 100kpb binned A[T>C]N and B[T>C]N counts. Thin orange line: A[T>C]/B[T>C] ratio, thick orange line: rolling average over 5 consecutive bins.

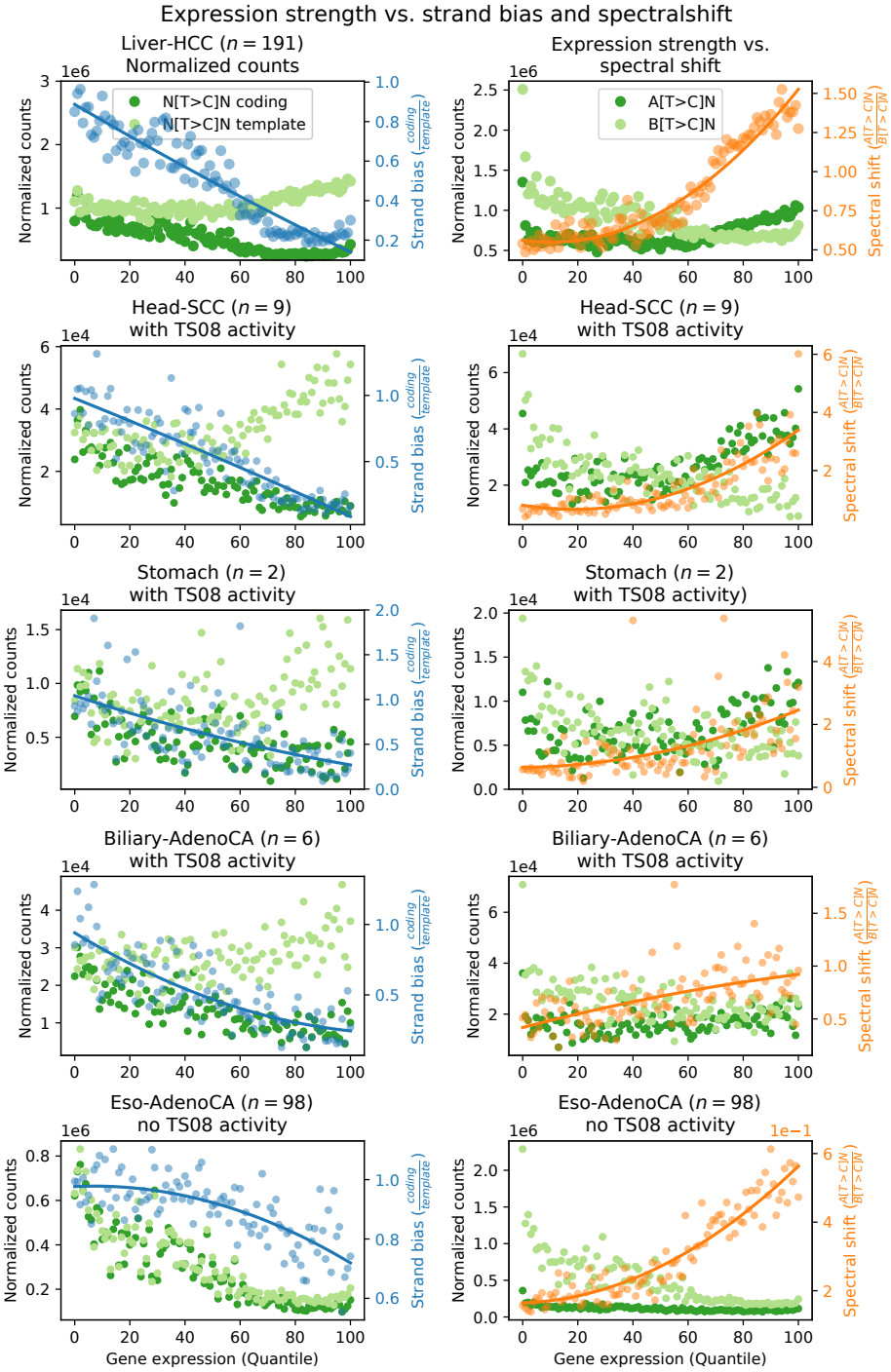


Fig. 3.13 **Transcription strand bias and spectral shift in samples from different cancers with TS07 and TS08 contributions.** Blue and orange lines correspond to quadratic fits of respective ratios.

quantiles (>50), but reverses on the template strand, producing more N[T>C]N mutations the higher the transcription, in line with previous reports of TAM (Haradhvala et al., 2016). Of note, this process mostly generated A[T>C]N mutations, in line with our signature inference. This effect is commonest in Liver-HCC samples, but is also found in Head-SCC, Stomach-AdenoCa and Biliary-AdenoCa (Fig. 3.13), showing that A[T>C]W TAM and N[T>C]N mutagenesis in quiet regions occur in a range of cancers and also normal Esophagus (Martincorena et al., 2018). In fact, it has been observed that SBS5, one of three widely active signatures, displays signs of potential contamination by SBS16/TS08, which may be more precisely resolved by the genomically informed TensorSignature analysis.

3.1.5 Replication- and DSBR-driven mutagenesis by APOBEC3A and APOBEC3B

In the following, we turn our focus to TS11-T[C>D]W;SV and TS12-T[C>D]W, which both share a base substitution spectrum attributed to APOBEC mutagenesis, but differ greatly with regard to their replicational strand bias, broader mutational composition, and clustering properties. While TS12 is dominated by SNVs (99 %) with strong replicational strand bias, SNVs in TS11 make up only 64 % of the overall spectrum and are highly clustered. The rest of the spectrum is mostly dominated by structural variants (Fig. 3.14, upper panel; Fig. C.53). This signature split reveals two independent triggers of APOBEC mutagenesis, which is thought to require single stranded DNA as a substrate, present either during lagging strand replication, or double strand break repair (DSBR). In the following, we will further assess the genomic properties of these two different modes of action.

Pooled base subtitution spectra spectra verify the genomic properties of TS11 and TS12

To verify the split, we pooled C>G and C>T variants from 30 and 15 samples with high TS11 and TS12 exposures, respectively (TS11 and TS12 contributions >10 % and 70 % respectively, Fig. 3.14, lower panel). We noticed that the spectrum in TS12-high samples was clearly dominated by T[C>D]N mutations, whereas the distribution in TS11-high samples was cross-contaminated by other mutational processes. However, assessment of replicational strand biases revealed that lagging strand mutations were twice as large as leading strand mutations in TS12-high samples, but not in TS11-high samples. Moreover, the proportion of clustered variants in TS12-high samples was much lower than in TS12-high in line with the signature inference (Fig. 3.14, lower panel).

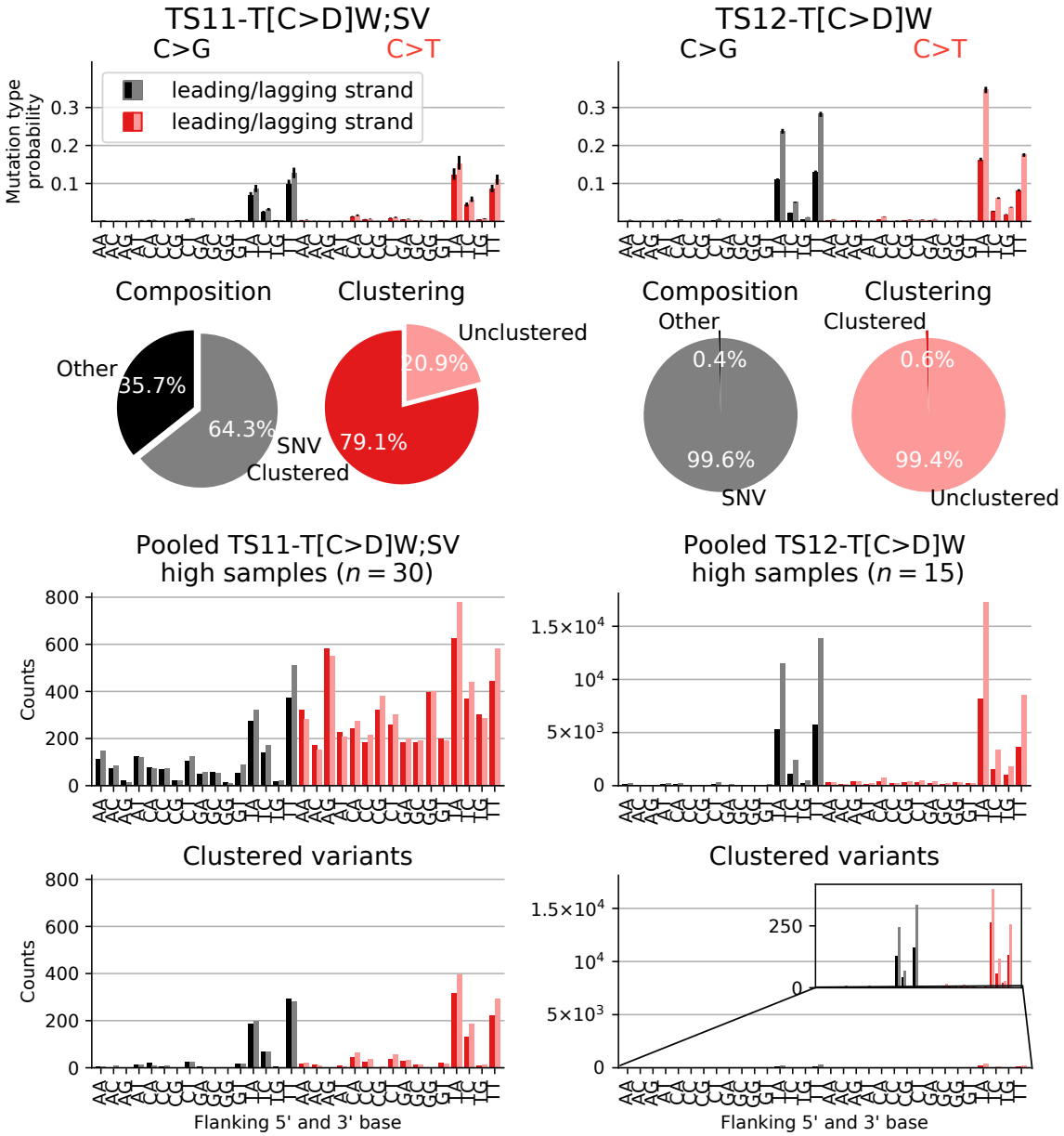


Fig. 3.14 **Diverging properties of TS11 and TS12.** Upper panel: C>G and C>T spectra of TS11 and TS12 for leading and lagging strand DNA. Pie charts underneath indicate percentages of clustered mutations and the contribution of other mutation types in TS11 and TS12. Lower panel: Observed unclustered (top) and clustered variants (bottom) in TS11 and TS12 high samples.

The association of SVs as well as genomic properties of TS11 indicate colocalisation of clustered mutations at sites of structural variation

The association of TS11 with structural variants suggests clustered APOBEC mutagenesis at sites of DNA double strand break events. This is confirmed by the spatial co-occurrence

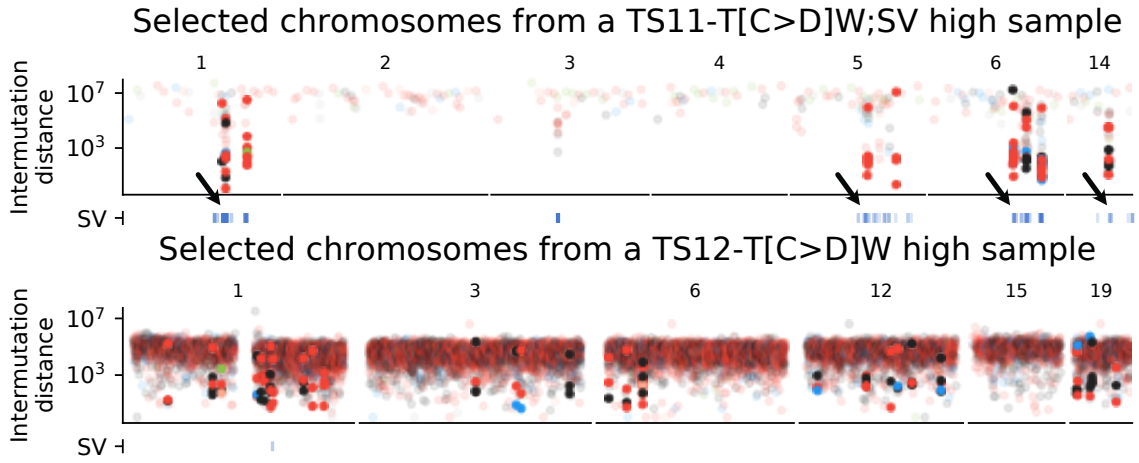


Fig. 3.15 **TS11 mutation cluster coincide with sites of structural variation.** Rainfall plots with SV annotations from a typical sample with high TS11 (top) and TS12 contributions (bottom).

of SVs and clustered mutations (a feature not directly measured by TensorSignatures; Fig. 3.15). Furthermore, SV-proximal clustered variants do not display a replicational strand bias, adding further weight to the notion that these arise in a DSBR-driven, replication-independent manner (Fig. B.8). Interestingly, SV-distal clusters displayed, on average, only a very weak replicational strand bias, indicating that the majority of these foci arose in a replication-independent fashion, presumably during successful DSBR, which did not create SVs.

TensorSignatures helps to tell apart differences in the distribution of clustered variants due to replication- and DSBR driven APOBEC mutagenesis

Next, we assessed whether differences exist in the characteristics of clustered variants, beyond the fact that these are much more frequent in DSBR driven mutagenesis. To this end, we pooled clustered variants from TS11/12-high samples and computed their size distribution, which revealed that the length of mutation clusters tend to be larger at SVs (Median 717 vs. 490bp, Fig. 3.16). This goes in line with the observation that clustered mutations at DSBRs tend to have more mutations per cluster (Median 5 vs. 4 variants; Fig. 3.16).

Differential size distributions of clustered variants suggest APOBEC isoform-specific activities

Differential size distributions of TS11/12 mutation clusters raise the question if an APOBEC subtype-specific activity is underlying their distinct genomic manifestation. Previous studies

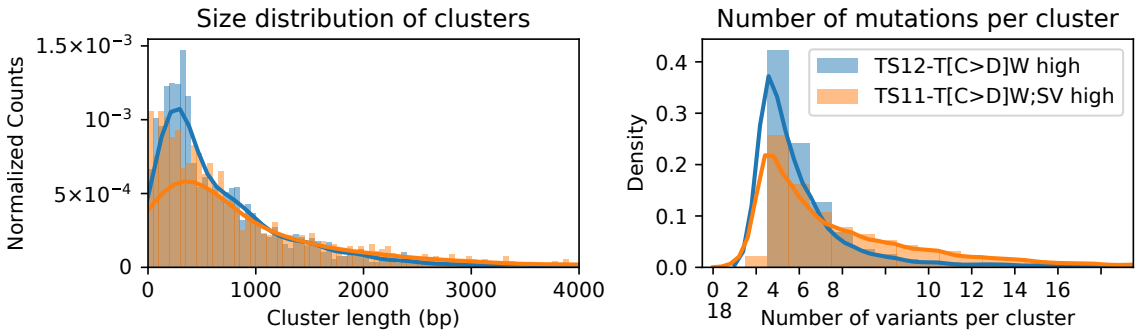


Fig. 3.16 **Size distribution of TS11 and TS12 mutation clusters.** Size distribution of mutation clusters (consecutive clustered mutations), and the distribution of number of variants per mutation cluster in TS11 and TS12 high samples respectively. Curves depict corresponding kernel density estimates.

linked the motifs YT[C>T]A and RT[C>T]A to APOBEC3A and APOBEC3B mutagenesis, respectively (Roberts et al., 2013). To test whether TS11 or TS12 clusters may be linked one or the other isoform, we extracted the pentanucleotide context at clustered T[C>T]A sites from samples with high TS11 or TS12 contributions. Clustered TS12 mutations comprise only a small fraction of purines, while this proportion increases to approximately 50 % in TS11 samples. These findings may indicate larger contributions of APOBEC3A and 3B in TS12 and TS11 samples, respectively (Fig. 3.17, B.9). To confirm this, we assessed samples that harboured a germline copy number polymorphism that effectively deletes APOBEC3B (Nik-Zainal et al., 2014). This analysis was unfortunately inconclusive, as the subset of PCAWG samples for which this annotation was available, did not show any activity for TS11.

Taken together, these results indicate that there are two distinct triggers of APOBEC mutagenesis, induced by DSBR or replication. Higher rates, longer stretches and larger proportions of RT[C>T]A APOBEC mutation clusters in the vicinity of SVs, as evidenced by TS11, suggests that DSBR leads to larger and possibly longer exposed stretches of single-stranded DNA, possibly due to APOBEC3B. Conversely, lower rates, shorter stretches and a high fraction of YT[C>T]A mutation clusters of TS12 in conjunction with a strong replicational strand bias indicate APOBEC3A mutagenesis during lagging strand synthesis, which is more processive than DSBR, allowing for fewer and shorter mutation clusters only.

3.1.6 Clustered somatic hypermutation at TSS and dispersed SHM

Two other TensorSignatures produced substantial amounts of clustered variants with, but different epigenomic localisation. TS13-N[C>K]H showed largest activities in lymphoid cancers and produced 60 % clustered variants (Fig. 3.5). The SNV spectrum resembles

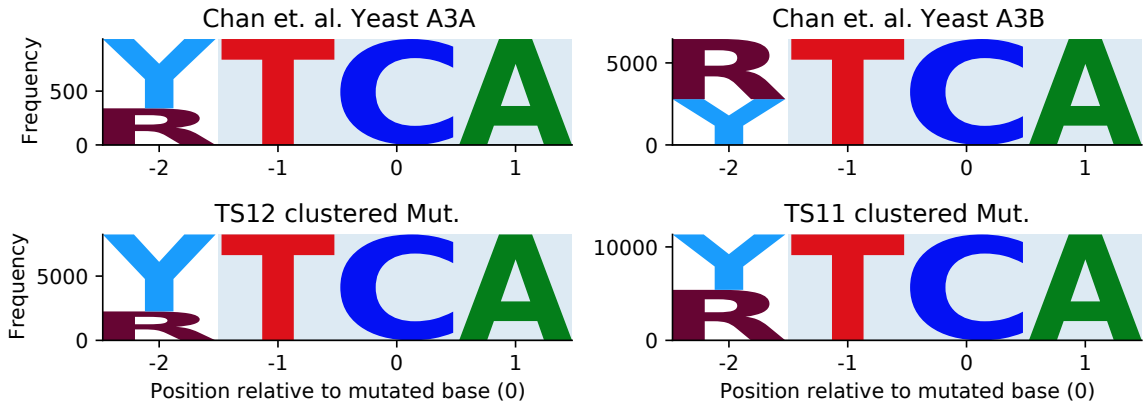


Fig. 3.17 **Higher order tetranucleotide motif logo plots at clustered TS11 and TS12 mutations indicate prevalent APOBEC3A and 3B mutagenesis.** Motif logo plots of the tetranucleotide context at mutated TCA sites in yeast cells exposed to APOBEC3A and 3B mutagenesis respectively (Chan et. al.), and similar motif logo plots extracted at clustered mutations from samples with high TS11 or TS12 exposures.

the c-AID signature reported previously (Kasar et al., 2015), suggesting an association with activation-induced cytidine deaminases (AID), which initiates somatic hypermutation in immunoglobulin genes of germinal centre B cells. Like its homolog APOBEC, AID deaminates cytosines within single stranded DNA, although it targets temporarily unwound DNA in actively transcribed genes, rather than lagging strand DNA or DSBRs (Muramatsu et al., 2000; Pham et al., 2003).

TS13 activity is strongly associated with transcription start sites, while mutations of TS14 disperse genome wide

TensorSignatures analysis reveals that TS13 activity is 9× and 8× enriched at active transcription start sites (TssA) and flanking transcription sites (TxFlnk, Fig. 3.5), respectively. To illustrate this, we pooled single base substitutions from Lymph-BHNL samples and identified mutational hotspots by counting mutations in 10 kb bins (Fig. 3.18). Inspection of hotspots confirmed that clustered mutations often fell accurately into genomic regions assigned as TssA (Fig. 3.18). The aggregated clustered mutation spectrum in TssA/TxFlnk regions across lymphoid neoplasms (Lymph-BNHL/CLL/NOS, $n = 202$) indeed showed high similarity to TS13, possibly with an even more pronounced rate of C>K (K=G or T) variants similar to SBS84 (Fig. 3.18, Alexandrov et al. (2019)). Conversely, the clustered mutational spectrum from all other epigenetic regions was characterised by a larger proportion of T>C and T>G mutations, similar to TS14-W[T>V]W, which only produces about 1 % clustered mutations

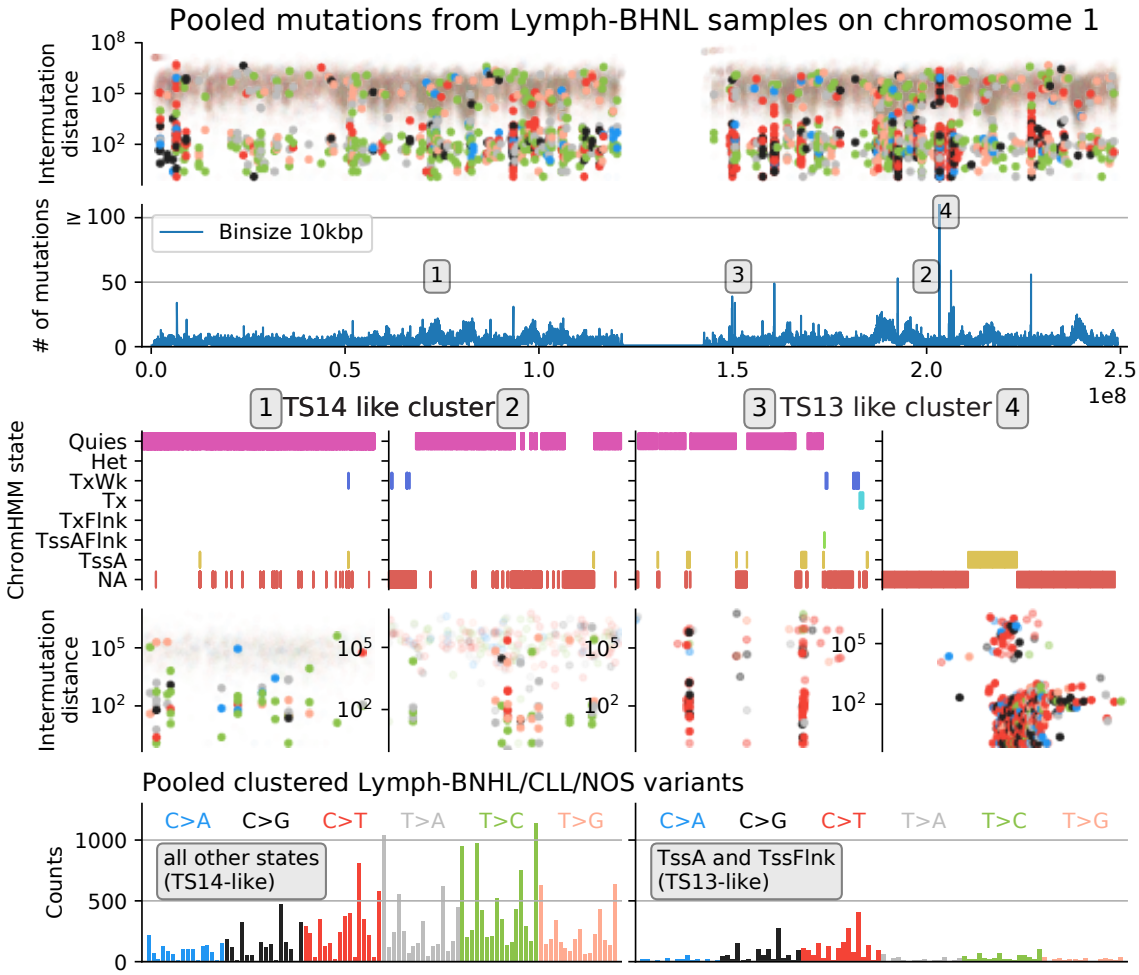


Fig. 3.18 **TS13 and TS14 mutation clusters occur at genomically distinct regions.** Upper panel: Rainfall plot of pooled variants from Lymph-BHNL samples on chromosome 1 (highlighted dots indicate clustered mutations). Binned (10 kb) SNV counts of chromosome 1. Numbers 1-4 indicate mutation hotspots. Middle panel: Consensus ChromHMM states and rainfall plots of mutation hotspots. Lower panel: Pooled clustered variants from PCAWG Lymph-BHNL/CLL/NOS samples from TssA or TxFlnk (TS13-like), and all other epigenetic states (TS14-like).

and closely resembles SBS9, attributed to Pol η-driven translesion synthesis (TLS) during somatic hypermutation.

TS13 mutation cluster are longer and contain more variants per cluster in comparison to TS14 mutation cluster

While TS13 and TS14 are strongly correlated ($R^2 = 0.88$, Fig. B.10), the diverging localisation pattern and SNV spectrum, characterised by higher rates of C>K mutations in

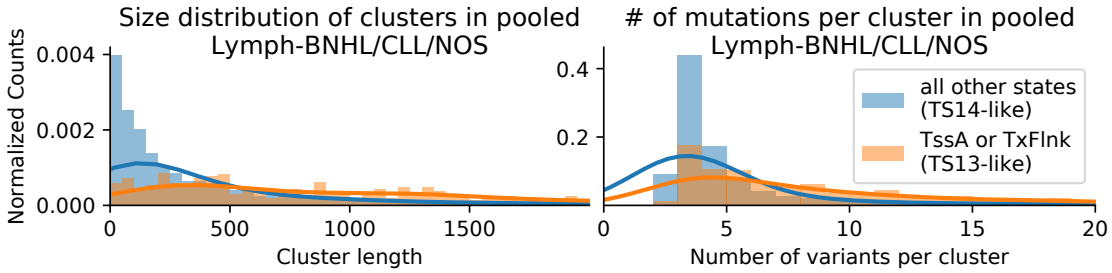


Fig. 3.19 **Size distribution and number of mutations in TS13 and TS14 mutation clusters.** Size distribution of mutation clusters (consecutive clustered mutations), and the distribution of number of variants per mutation cluster in TS13 and TS14 high samples respectively.

TS13, indicates that a related, but different mutational process drives TSS hypermutation, seemingly linked to AID. The differential mechanism behind TS13 also manifests as longer clusters (Median: 1,068 vs. 183bp), which contain more variants per cluster (Median: 8 vs. 3 mutations) in comparison to TS14 (Fig. 3.18).

TS14 is found in a broad range of cancers and is predominantly active in late replicating region; TS13, on the other hand, in early replicating regions

As a further distinction, the weakly clustered TLS signature TS14 can be found in more than 15 cancer types, suggesting a broad involvement of this mutagenic process in resolving endogenous and exogenous DNA alterations (Supek and Lehner, 2017). Pol η has also been described to compete with lagging strand DNA synthesis (Kreisel et al., 2019), which is further corroborated by the fact that TS14 displays a mild replicational strand bias (RSB = 0.9; Fig. 3.5). Interestingly, TS14 is found to be predominantly active in regions without replication orientation ($a_{\mathrm{RS}} = 0.7$), which are usually far from the origin of replication (Fig. 3.5). Conversely, TS13 is mostly found in oriented, early replicating regions, but does not display a measurable replicational strand bias (Fig. 3.5), indicating different modes of activation.

The HMF cohort contains a third mutational signature of somatic hypermutation

Finally, a third mutational signature of somatic hypermutation, TS30 (see next section Fig. 3.22), was found in lymphoid and other cancers of the HMF cohort. This signature displayed a large proportion of clustered mutations and an enrichment in early replicating regions similar to TS13, combined with an SNV spectrum that was closer to TS14 (Cosine distance 0.13 vs. 0.25), suggesting that TS30 may represent a combination of TS13 and TS14.

3.2 Validating tensor signatures in the HMF cohort

The aforementioned observations were replicated in a fully independent second cohort of whole genomes from the Hartwig Medical Foundation with 3,824 samples from 31 cancers encompassing 95,531,862 SNVs, 1,628,116 MNVs, 9,228,261 deletions, 5,408,915 insertions and 1,001,433 structural variants (Priestley et al., 2018).

3.2.1 Applying TensorSignatures to the genomes of the HMF cohort produced 27 tensor signatures

Applying TensorSignatures to this data set produced 27 tensor signatures (Fig. 3.20, 3.22, 3.23, and B.11). Of these 10 closely resembled (cosine distance < 0.2) signatures of the discovery analysis with closely matching genomic activity coefficients (Fig. 3.21, B.12). These include the signatures of spontaneous deamination TS01, the two signatures of UV mutagenesis TS05/06, SV-associated APOBEC mutagenesis TS11, as well as signatures of MMRD TS16, POLE^{exo} mutations TS17, as well as MUTYH deficiency TS18, HRD TS19 and TS20.

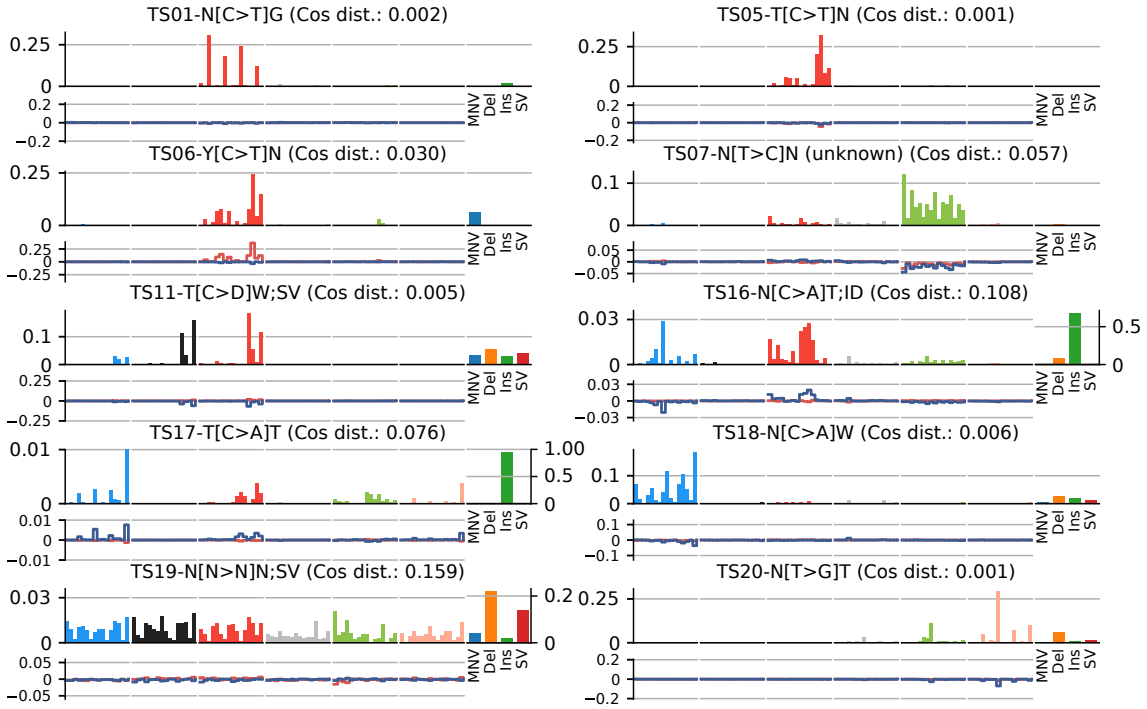


Fig. 3.20 **Validated HMF signatures.** Validated tensor signatures with high similarity (indicated as cosine distance) to the mutational processes extracted in our discovery analysis using PCAWG data (representation analogous to Fig. 3.4).

Fig. 3.21 **Tensorfactors and exposures of the HMF cohort.**. Extracted tensor factors, exposures and summed squared errors of tensor factors from the discovery and validation analysis.

A further 7 signatures seemingly constitute splits of tensor signatures from the PCAWG cohort (Fig. 3.22). A complex three-way split appeared to occur for TS03 and TS04, which were found in a broad range of cancer types. One of the derivative signatures resembles the mutation spectrum of SBS8 from the COSMIC catalogue, however without measurable transcriptional strand bias. A second derived signature is similar to SBS39; our analysis reveals replication strand bias for C>G variants and a potentially wider range of cancer types for both signatures. Further, signature TS12, resembling replication associated APOBEC mutagenesis, split into two signatures with base substitution spectra similar to SBS2 (C>T) and SBS13 (C>G), but preserving the strong replication strand bias. Lastly, a split of TS10, likely attributed to mutagens included in tobacco smoke, was observed.

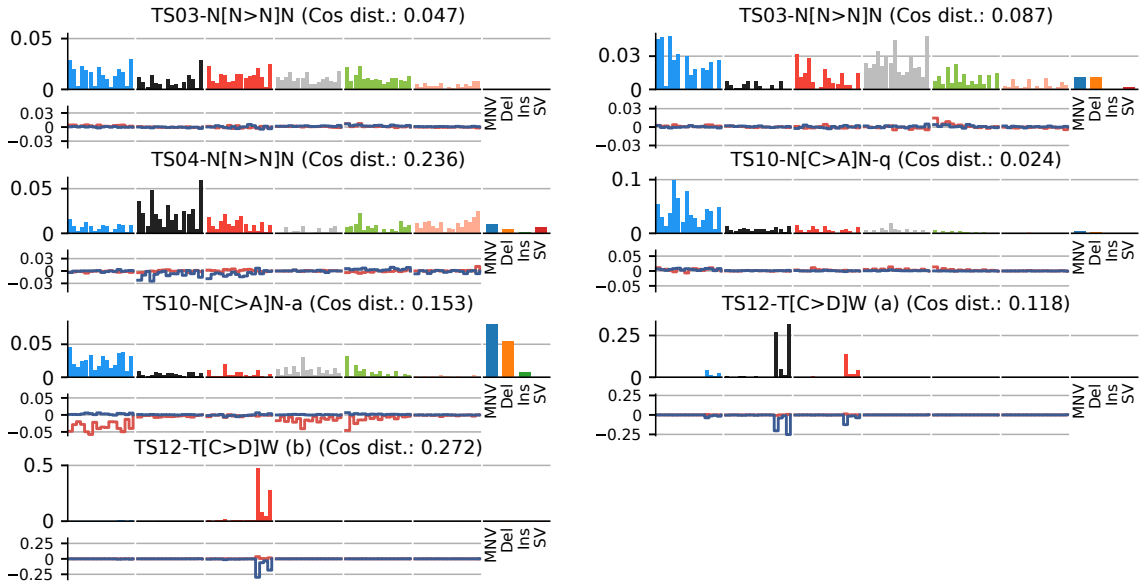


Fig. 3.22 **Split HMF signatures.** TensorSignature splits that seemingly represent derivatives of tensor signature TS03, TS04, TS10 and TS12. Single base substitution spectra and a summarised representation of other mutation types, as well as replication and transcription biases (representation analogous to Fig. 3.4).

Finally a set of 10 novel signatures without close match to those in the PCAWG cohort was found (Fig. 3.23). This includes five spectra linked to cancer therapies, illustrating the additional insights on preceding therapies provided by the HMF metastatic cancer cohort. TS21 is characteristic of treatment with the methylating agent temozolomide (SBS11); the observed transcriptional strand bias reflects a higher rate of G>A mutations on the coding strand (equivalent to higher rates of C>T on the template strand), consistent with methyl guanine being removed by TC-NER in the absence of MGMT. TS22 and TS23 have been previously associated with cisplatin (termed E-SBS21 and E-SBS14, Christensen

et al. (2019); Pich et al. (2019)). While both signatures exhibit mild transcriptional strand biases, only TS23 shows a strong association with MNVs going in line with the propensity of cisplatin/oxaliplatin to form intrastrand DNA adducts (Fig. 3.23). TS24 displays the characteristics of treatment with 5-FU, which inhibits thymine synthesis and has been proposed to be mutagenic via genomic fluorouracil incorporation (Christensen et al., 2019). TS28, with similarity to SBS41, was only found in two samples, possibly due to treatment with the experimental drug SYD985, which consists of a duocarmycin-based HER2-targeting antibody-drug conjugate (Priestley et al., 2018).

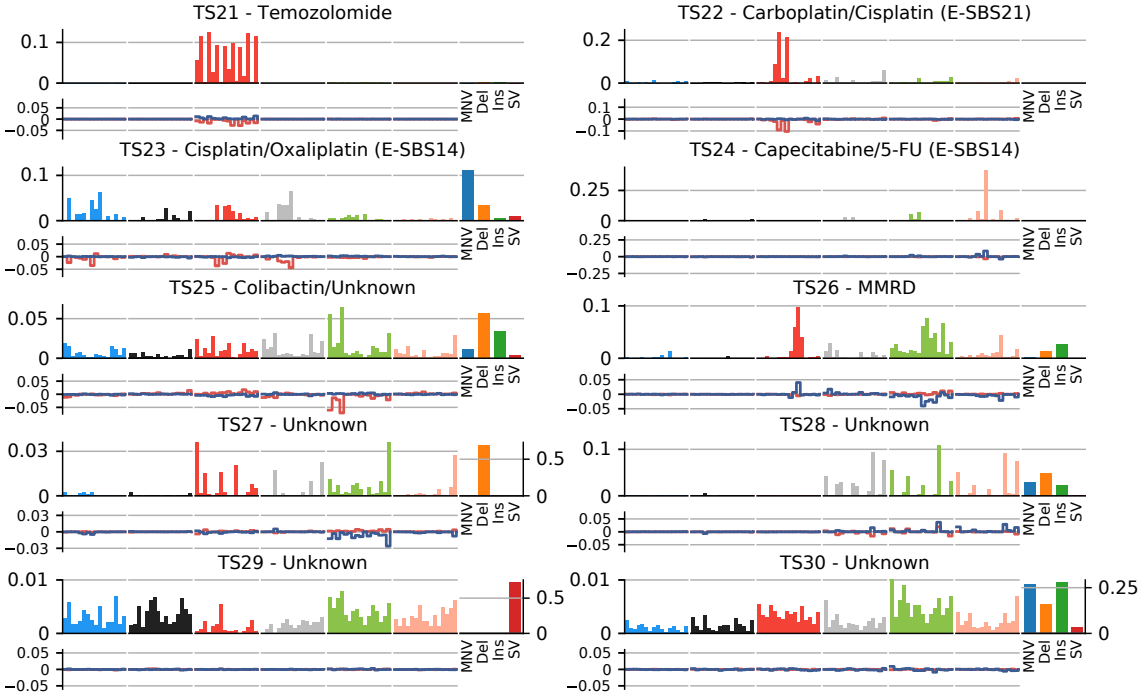


Fig. 3.23 **New HMF signatures.** Novel tensor signatures of the HMF cohort. Single base substitution spectra and a summarised representation of other mutation types, as well as replication and transcription biases (representation analogous to Fig. 3.4).

Further, TensorSignatures detected a signature of colibactin, TS25, which has been previously characterised (Pich et al., 2019; Pleguezuelos-Manzano et al., 2020). TS25 displays contributions of MNVs and short indels, activity in active genomic regions and concomitant transcriptional strand bias of T>C mutations (Fig. 3.23, 3.21). TS26's indels and similarity to SBS15 suggests an association with MMRD; TS27 has an unknown aetiology and displays strong replicational strand bias. The large proportion of structural variants and the flat SNV spectrum of TS29 may represent non-specific mutagenesis at SVs. TS30 was found in lymphoid and other cancers and had a high proportion of clustered mutations, similar, but not identical to TS14 (Fig. 3.21).

3.2.2 TC-NER changes the mutation spectrum of tobacco-assocatiated mutations

A similar split of an exogenous mutational signature into quiet and active chromatin was observed in lung cancers (pooled squamous cell and adeno-carcinoma) of the HMF cohort where TS10 splits into two signatures, HMF TS10-q, which shows largest activity in heterochromatin, while HMF TS10-a is enriched in actively transcribed regions, and exhibits a strong transcriptional strand bias with lower rates of C>A changes on the coding strand, equivalent to G>T transversions on the template strand (Fig. 3.22, 3.21, 3.24). This strand bias has been attributed to TC-NER removing benzo[a]pyrene derived adducts on guanines from the template strand (Pleasance et al., 2010b).

The emergence of two mutational signatures indicates that this repair process also changes the mutation spectrum. The suggested split is also evident in pooled mutations from HMF lung cancers in quiescent and active genomic regions, respectively, revealing that predicted spectra coincide with corresponding tensor signatures HMF TS 10-q and TS10-a (Fig. 3.24).

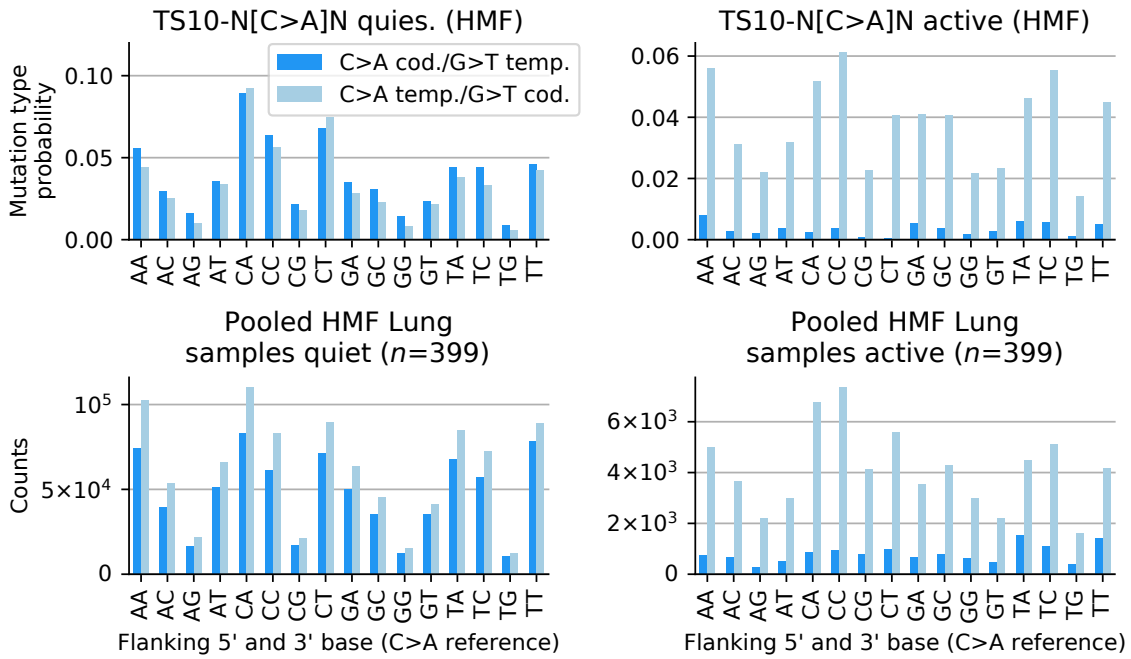


Fig. 3.24 **TC-NER changes the mutation spectrum of tobacco-assocatiated mutations.** C>A mutation type probabilities of HMF TS10-q and TS10-a for coding and template strand DNA.

The C>A (G>T) mutation spectrum observed in quiescent regions, extracted by HMF TS10-q, displays highest rates of mutations in a CCN (NGG) context (Fig. 5a). Interestingly,

the same pattern is also observed in actively transcribed regions for C>A on the template strand, equivalent to G>T mutations on the coding strand. This is in contrast to the C>A coding strand pattern, and HMF TS10-a, for which this difference is largely eroded. These observations reflect how TC-NER removes genotoxic guanine adducts from the template strand, which leads to lower mutation rates and also a more homogeneous base context of G>T mutations. The differentential mutation spectrum indicates that either the efficiency of TC-NER – or the mutagenicity of residual genomic alterations – differs depending on the base context, analogous to observations in UV-induced mutagenesis. The result being that the mutation types and rates caused by tobacco-associated carcinogens differ between coding and template strand in transcribed regions and also to different mutation spectra in quiescent and active genomic regions.

3.3 Further tensor signatures in cancer and normal tissues

During my PhD, I applied TensorSignatures to a variety of datasets. These included studies with the aim to delineate the genomic effects of proofreading deficiencies in replicative DNA polymerases (Sec. 3.3.1) and mismatch repair deficiency (Sec. 3.3.2), an analysis of the mutational landscape in oesophageal adenocarcinoma (Sec. 3.3.3) and human bladder urothelium (Sec. 3.3.4), as well as a comparative study with the goal to understand the mutational processes in normal and tumorous samples from various tissues (Sec. 3.3.5). Following sections briefly describe my contributions to these analyses.

3.3.1 Elevated somatic mutation burdens in normal human cells due to defective DNA polymerases

Replicative DNA polymerases Pol ε and Pol δ are characterised by remarkably low base substitution error rates, which are partly due to their intrinsic proofreading activities, enabling them to detect and remove wrongly incorporated bases during DNA replication (Sec. 1.2.3). However, germline mutations in the *POLE* and *POLD1* exonuclease domain may cause polymerase proofreading associated polyposis, which manifests as early-onset cancer, especially in fast dividing epithelia from tissues such as the colon, rectum and endometrium. An appealing hypothesis for the aetiology of this syndrome are elevated mutation rates due to the defects in the aforementioned enzymes. To understand the consequences of proofreading deficiencies in Pol ε and Pol δ, this study sequenced normal tissues from individuals with germline exonuclease mutations in *POLE* (L424V, $n = 8$) and *POLD1* (S478N, $n = 4$; L474P, $n = 1$; D316N, $n = 1$). Here, I analysed the somatic mutational catalogue of 211

samples, which represent branches in reconstructed phylogenetic trees of fourteen individuals, comprising a total of 1,971,246 SNVs, 2,594 MNVs, 26,869 deletions and 135,452 insertions.

Contributions

This chapter is based on data from the bioarxiv manuscript "Elevated somatic mutation burdens in normal human cells due to defective DNA polymerases" by Philip S. Robinson, Tim H.H. Coorens, Claire Palles, Emily Mitchell, Federico Abascal, Sigurgeir Olafsson, Bernard Lee, Andrew R.J. Lawson, Henry Lee-Six, Luiza Moore, Mathijs A. Sanders, James Hewinson, Lynn Martin, Claudia M.A. Pinna, Sara Galvotti, Peter J. Campbell, Inigo Martincorena, Ian Tomlinson, Michael R. Stratton. H.V. conducted the mutational signature analysis and produced the figures presented in this section. The authors acknowledged my efforts.

The mutational signatures of proofreading deficient replicative DNA polymerases

The TensorSignatures analysis revealed eight mutational processes (Fig. D.1), four of which closely resemble previously presented tensor signatures, as well as additional four mutational processes that seemingly represent the consequences of proofreading deficiencies. Former include TS01, indicative for spontaneous deamination of 5-meC, two cancer treatment associated signatures TS23 and TS24 (Pich et al., 2019), as well as a TS25-like signature, likely due to colibactin exposure (Pleguezuelos-Manzano et al., 2020). The remaining four mutational processes are characterised by sharp C>A peaks at TT contexts and strong replicational strand biases towards leading (TS17 and TS17-a) and lagging strand DNA (TS-POLD1 (Ins) and TS-POLD1), consistent with defects in the exonuclease domain of Pol ε and Pol δ, respectively (Fig. 3.25, D.2, D.3, D.4, and D.5).

Closer inspection of TS17 reveals low contributions of insertions and a decreased signature activity in transcribed epigenetic states (Tx, TxWk), while TS17-a exhibits a characteristic C>T pattern, closely resembling COSMIC SBS10b, and a slight enrichment in transcribed regions (Fig. 3.25). To verify the epigenetic activation pattern of both signatures, I pooled mutations from *POLE* L475V samples in transcribed and heterochromatic regions, revealing that the C>A/C>T ratio decreases from 1.2 in heterochromatic to 0.6 in transcribed regions, going in line with a higher activity of TS17-a in transcribed regions (Fig. D.6). A similar pattern was observed for both POLD1 tensor signatures, albeit the contribution of insertions in TS-POLD1 (Ins) makes up a substantially larger fraction of 31 %. Inspection of pooled single base substitutions in *POLD1* S478N samples indicated a more balanced pattern of

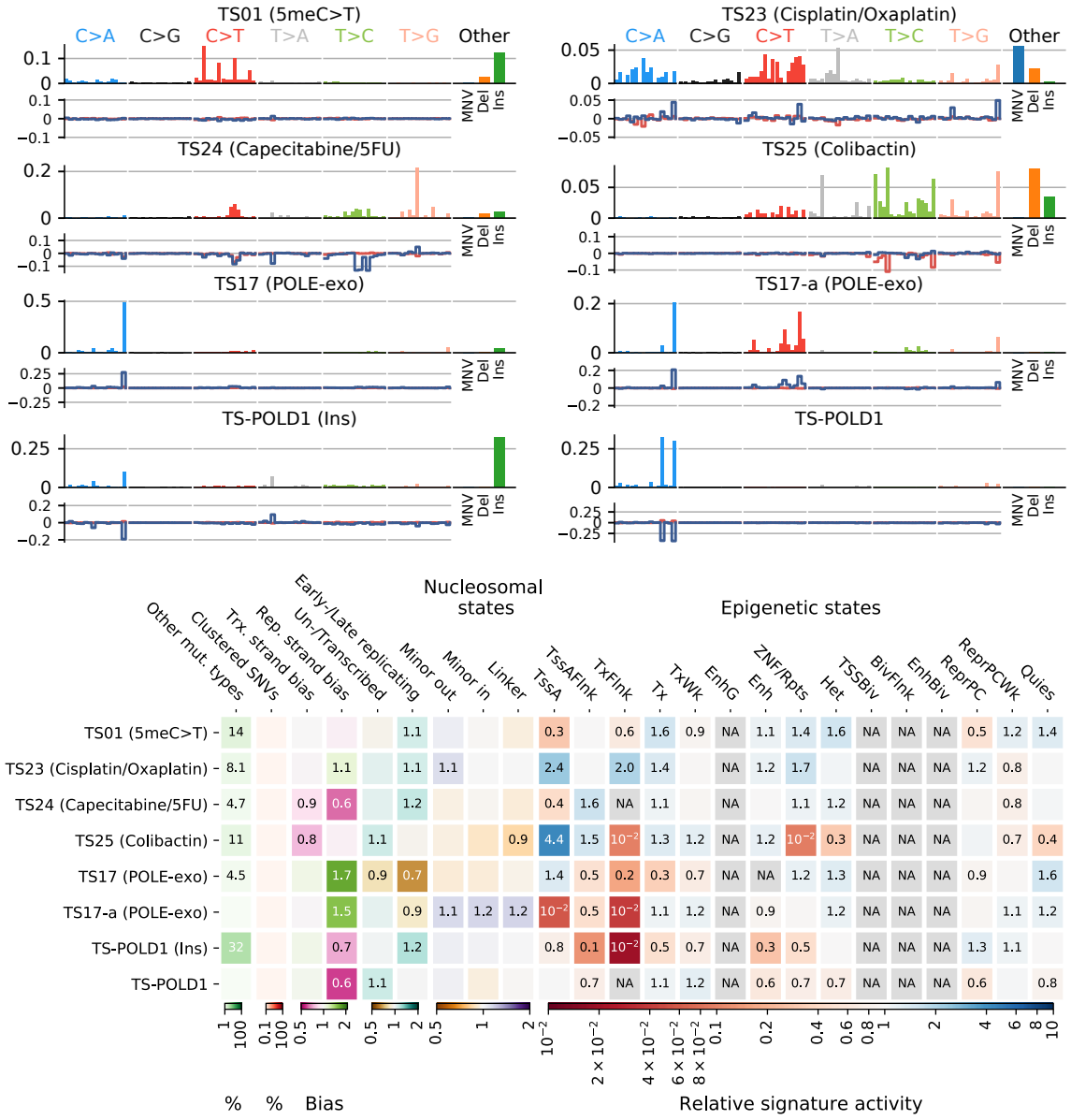


Fig. 3.25 **Tensor signatures and accompanying tensor factors of proofreading deficient DNA polymerases (Robinson et al., 2020).** Upper panel: single base substitution spectra and a summarised representation of other mutation types, as well as replication and transcription biases (representation analogous to Fig. 3.4). Lower panel: accompanying tensor factors (representation analogous to Fig. 3.5).

C>A mutations in TT and TA context in transcribed regions, consistent with higher activity of TS-POLD1, as predicted by the inference (Fig. D.6).

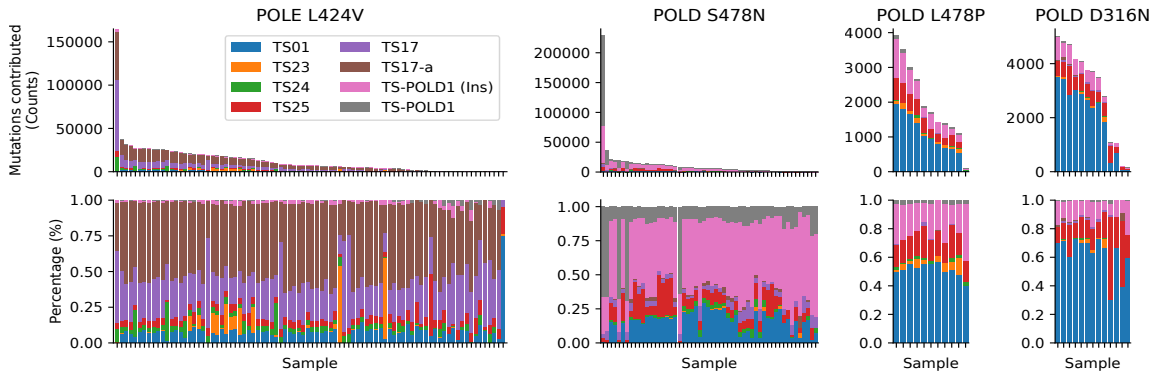


Fig. 3.26 **The signature composition of samples with defects in the proofreading domain of replicative DNA polymerases (Robinson et al., 2020).** Absolute and relative signature contributions, each bar represents a sample.

POLD1 L478P and D316N samples exhibit TS-POLD1 (Ins) mutations but lack TS-POLD1 activity

The attribution of TS17 and TS17-a to *POLE* defects, as well as TS-POLD1 (Ins) and TS-POLD1 to *POLD1* deficiencies is corroborated by analysing the signature composition of respective samples (Fig. 3.26). Samples harbouring the L472V mutation in the exonuclease domain of Pol ε are largely composed of TS17-a and TS17 mutations, while most *POLD1* S478N samples are dominated by TS-POLD1 (Ins) mutations, and a few samples with substantial contributions of TS-POLD1. On the other hand, *POLD1* L478P and D316N exhibit milder contributions of TS-POLD1 (Ins), and almost no TS-POLD1 contributions, indicating that these mutations cause a less severe phenotype in comparison to the S478N mutation.

3.3.2 The mutational signatures of DNA mismatch repair deficiencies

The DNA mismatch repair pathway is crucial to ensure high fidelity DNA replication, and is mediated in humans by the orchestration of the proteins Msh2, Msh6, Mlh1 and Pms2. DNA mismatches occur when DNA polymerases insert wrong bases opposite to template DNA and fail to excise them via their proofreading activities, or at sites of small insertions and deletions. Shortly after DNA replication, the Msh2:Msh6 heterodimer MutS scans newly synthesised DNA for mismatches, and recruits upon lesion detection a second heterodimer MutL composed of Mlh1 and Pms2, that nicks the DNA segment containing the mismatch, thus preparing it for excision and subsequent repair (Sec. 1.4.2). This study aimed to characterise the genomic effects of germline mutations in *MSH6*, *PMS2* and *MLH1* in a cohort of 215 samples, which represent branches in reconstructed phylogenetic trees of 15

individuals. The dataset comprised 1,253,549 SNVs, 4,200 MNVs, 359,471 deletions and 252,982 insertions.

Contributions

This sections contains preliminary results obtained in collaboration with Mathijs A. Sanders and contains the results of a manuscript in preparation. H.V. conducted the mutational signature analysis and produced the figures presented in this section.

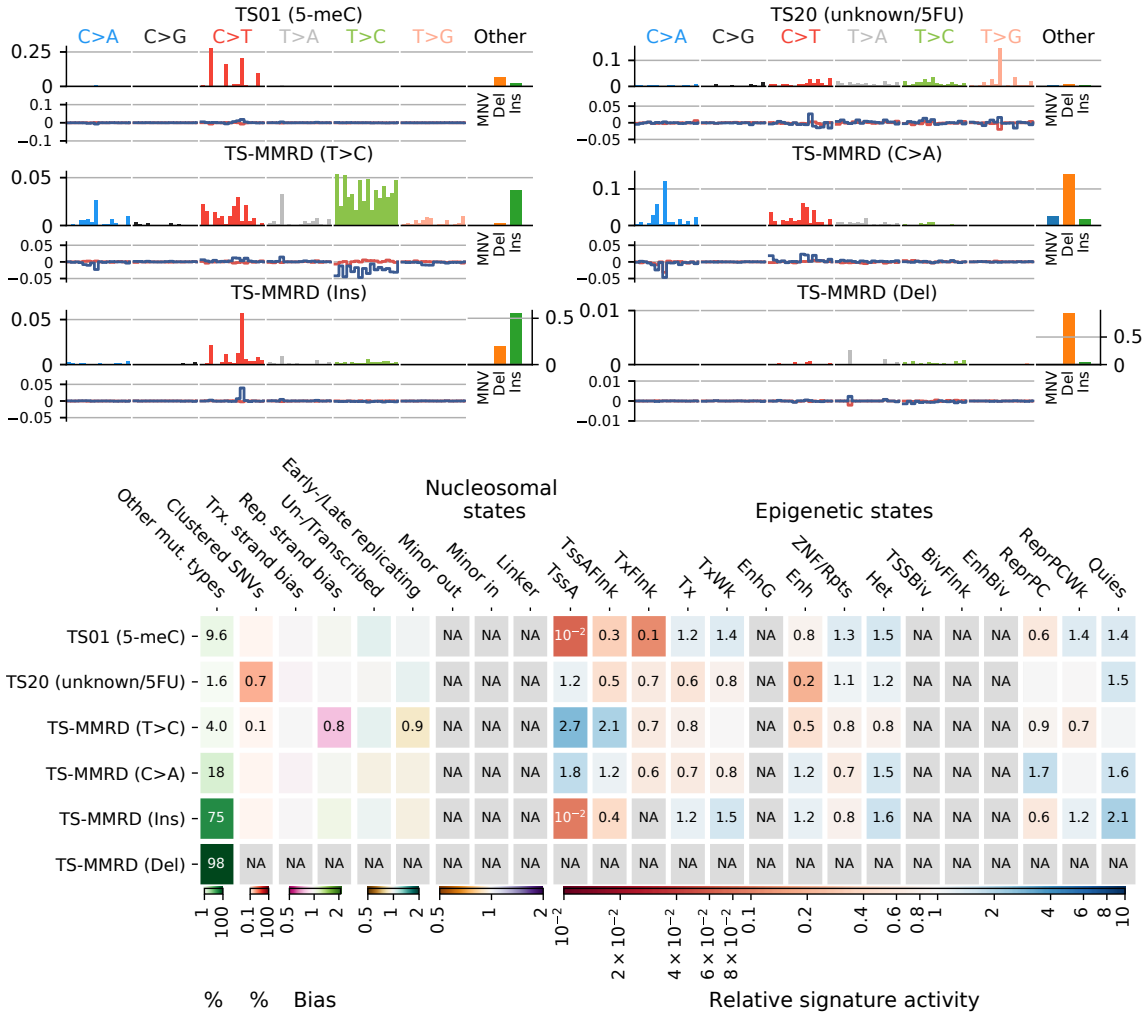


Fig. 3.27 **Tensor signatures and accompanying tensor factors of mismatch repair deficiencies.** Upper panel: single base substitution spectra and a summarised representation of other mutation types, as well as replication and transcription biases (representation analogous to Fig. 3.4). Lower panel: accompanying tensor factors (representation analogous to Fig. 3.5).

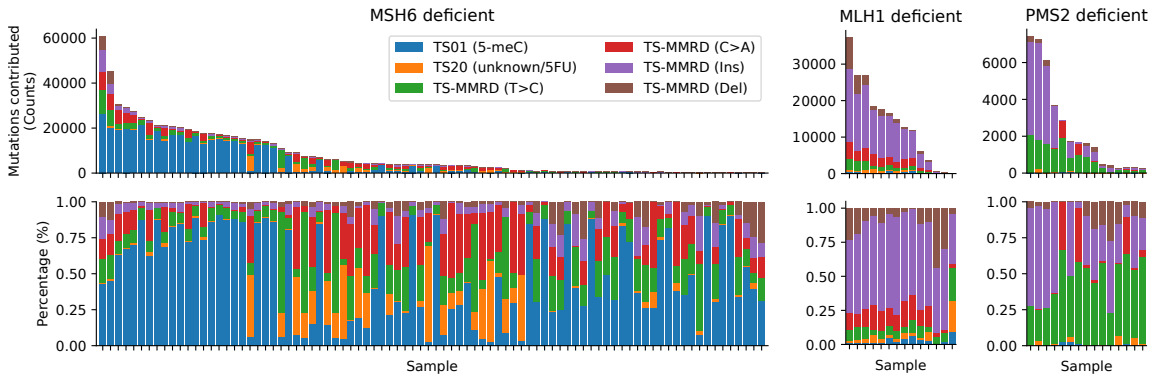


Fig. 3.28 **The mutational composition of samples with defects in the mismatch repair pathway.** Absolute and relative signature contributions, each bar represents a sample.

The analysis of 215 MMRD deficient samples revealed six tensor signatures

The analysis revealed six mutational processes (Fig. D.7), comprising TS01 and TS20, representative for spontaneous deamination of 5-meC and an unknown mutational process (possibly 5-FU treatment), respectively, as well as four additional signatures, which are likely to depict the mutational imprints of mismatch repair deficiencies (Fig. 3.27). TS-MMRD (T>C) base substitution spectrum is characterised by T>C mutations, smaller contributions of insertions, and a slight replicational strand bias towards lagging strand DNA (Fig. D.8). TS-MMRD (C>A) exhibits a sharp C>A peak at a CT context, and comprises roughly 20 % of other mutation types (Fig. D.9). Finally, there are two mutational processes, TS-MMRD (Ins) and TS-MMRD (Del), which are mainly consist of insertions and deletions, respectively, although the former signature also features a distinct C>T pattern with a peak at GG contexts (Fig. D.10, D.11).

MSH6 deficient samples fail to repair demethylated 5me-C, while defects in the MMR endonuclease lead to the accumulation of insertions

The decomposition of active mutational processes in *MSH6* deficient samples revealed TS01 as a major contributor to the mutational landscape, suggesting that the MutS protein complex fails to detect T:G mispairs, which result from spontaneous demethylation of 5-meC (Sec. 1.3.1). Moreover, *MSH6* deficient samples showed elevated activities of TS-MMRD (T>C) and TS-MMRD (C>A), but harboured only smaller contributions of TS-MMRD (Ins) and TS-MMRD (Del), indicating that the recognition of indels is less impaired. In contrast, *MLH1* and *PMS2* deficient samples showed strong contributions of the insertion signature TS-MMRD (Ins), suggesting that the endonuclease activity of MutL is crucial to deplete this mutation type from the genome (Fig. 3.28).

3.3.3 The mutational landscape of oesophageal adenomacarcinoma

Oesophageal cancer is one of the most prevalent form of malignant tumours in western countries, and ranks as the sixth most common cause of cancer-related death worldwide (Frankell et al., 2019). Particularly, oesophageal adenoma carcinomas (OACs) are highly aggressive, clinically late diagnosed and often highly resistant to chemo therapy. Also, OACs exhibit very high mutation rates, as well as marked chromosomal instability, and are thus classified as C-type neoplasm, which are predominantly caused by SVs rather than mutations (Frankell et al., 2019). The OCCAMS study aimed to characterise genomic biomarkers in OAC by analysing the genomes from patients recruited across the UK. Here, I applied TensorSignatures to a dataset of 383 OAC samples with variable chemo therapy status, comprising 12,877,577 SNVs, 77,034 MNVs, 623,030 deletions, 327,308 insertions and 106,599 SVs.

The analysis of the OCCAMS dataset revealed 15 tensor signatures

The analysis of the OCCAMS dataset revealed 15 mutational processes, indicative for the highly heterogenous mutational landscape of OACs. Ten of these mutational processes closely resemble previously described signatures: TS01, descriptive for spontaneous demethylation of 5-meC; an TS08-like signature, possibly due to transcription coupled mutagenesis; TS12, indicative of APOBEC mutagenesis on lagging strand DNA; TS15, most likely due to MMRD; TS18, indicative of defects in the BER pathway; TS11, implying failure of homologous recombination; two signatures TS20-early and TS20-late, possibly associated with 5-FU treatment, and whose split shall be discussed later on; TS23, indicative of cancer treatment with cisplatin; and a TS25-like signature, which is associated with the exposure to colibactin. In addition, the tensor decomposition gave rise to five novel signatures. These include TS-OAC (flat), which is slightly clustered and exhibits a relatively flat spectrum, as well as TS-OAC (C>T), whose single base substitutions remind of COSMIC SBS30, a signature that has been associated with defects in the *NHTL1* gene of the BER pathway (Sec. 1.6). The remaining three signatures TS-OAC (Indel), TS-OAC (Del) and TS-OAC (SV) were associated with other mutation types and displayed rather cryptic SBS spectra.

The tensor factorisation revealed a split of TS20, which is likely to be driven by differential activity in early and late replicating regions (Fig. D.13). In agreement with this hypothesis, TS20-early exhibits a slightly stronger replicational strand bias in comparison to TS20-late, and is detectable in epigenetically active genomic regions (Fig. D.13), which tend to colocalise at early firing ORIs.

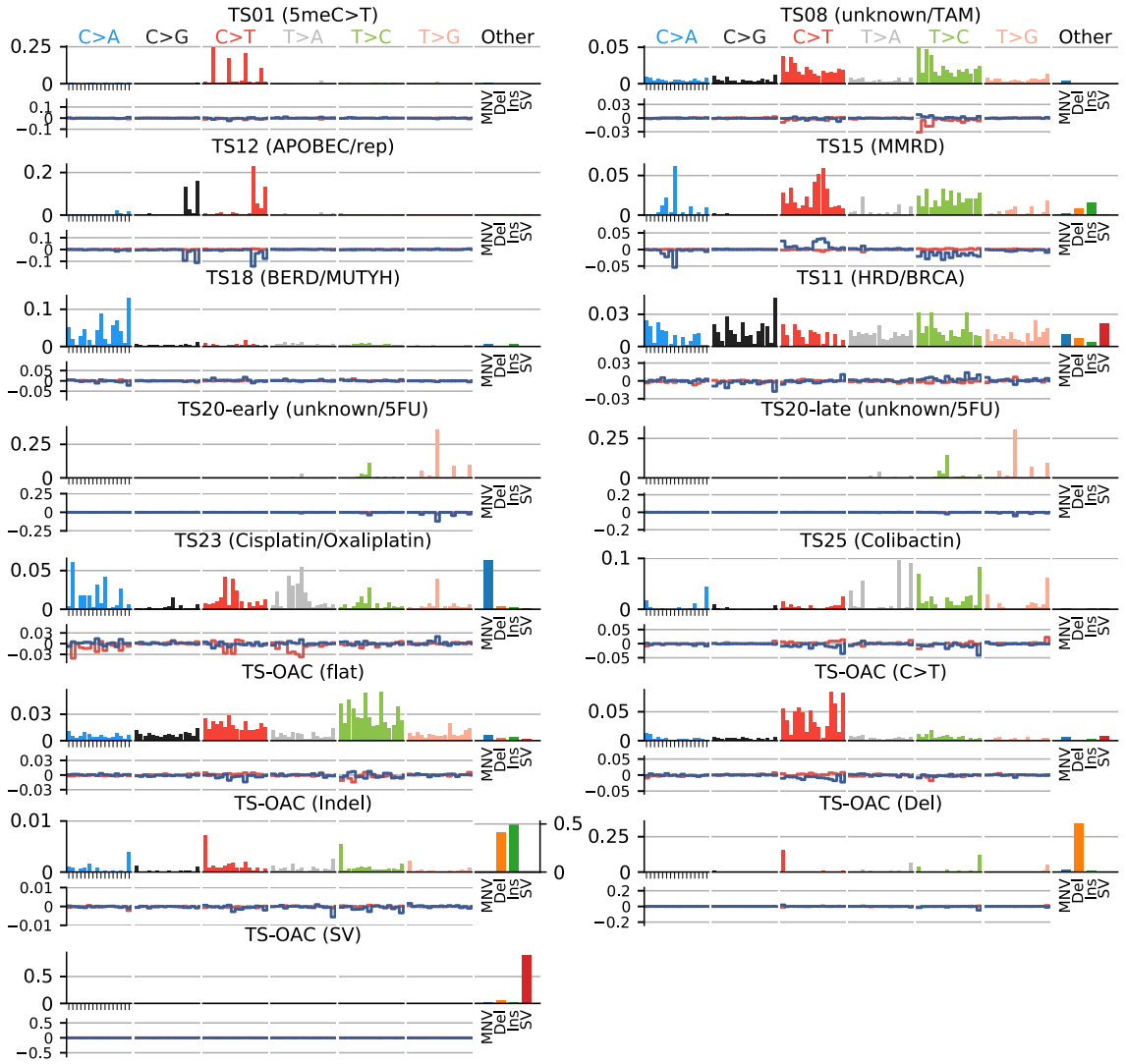


Fig. 3.29 **The tensor signatures of oesophageal adenomacarcinomas.** The representation of tensor signatures analogous to Fig. 3.4.

The mutational composition of chemotherapy naive and treated OACs

Next, I analysed the mutational composition of chemotherapy naive and treated OACs (Fig. 3.30), which both exhibit a very heterogenous mutational composition, often indicating substantial contributions from four or more mutational processes. The great majority of samples show high activity of TS18, TS20-early and TS20-late, as well as to lower extend contributions from TS25 and TS-OAC (C>T). Samples with high mutational burden show strong activities of TS08, TS15 and TS-OAC (Del). The co-occurrence of TS-OAC (Del) with MMRD signature TS15 suggests microsatellite instability as a putative cause.

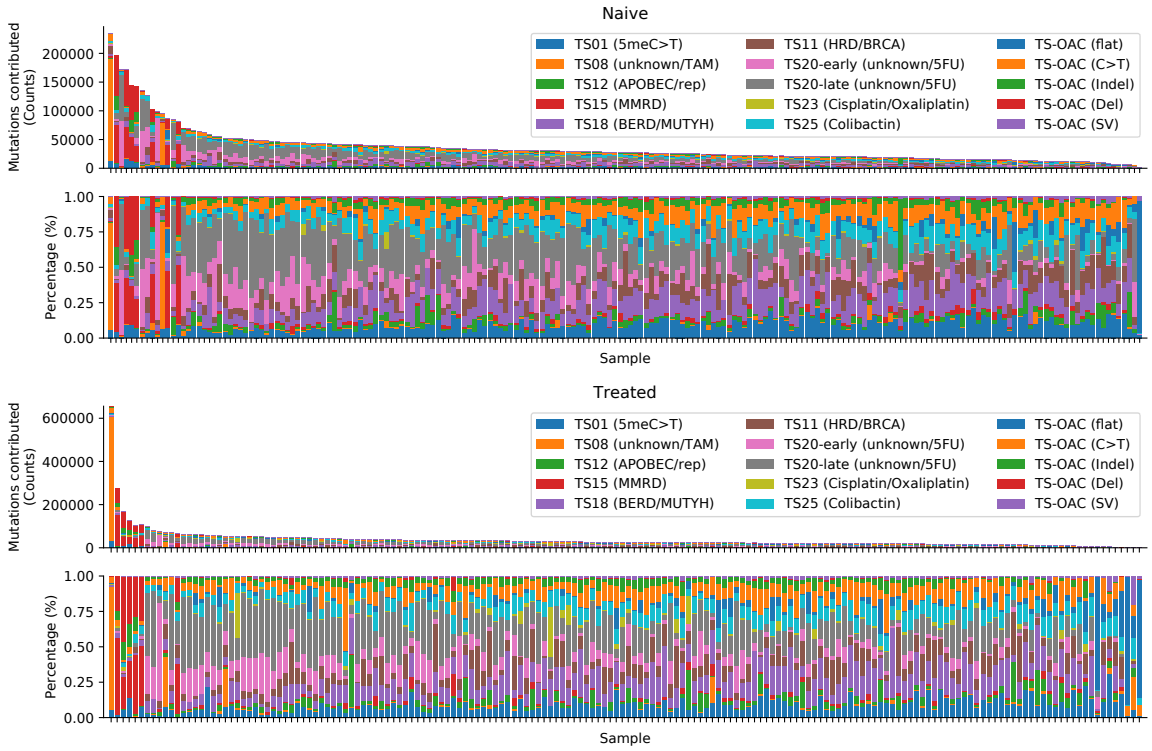


Fig. 3.30 **The signature composition of chemo therapy naive and treated oesophageal adenocarcinoma.** Absolute and relative signature contributions, each bar represents a sample.

3.3.4 Extensive heterogeneity in somatic mutation and selection in the human bladder

Bladder urothelium is constantly exposed to urine, which may contain mutagenic molecules such as aromatic amines from tobacco smoke or aristocholic acid from herbal remedies. Although the tissue is known to quickly regenerate upon injuries, bladder urothelium is generally considered as a slowly dividing epithelium. This circumstance makes it surprising that cancers arising from urothelium represent one of the neoplasms with highest mutation rates of all major cancer types. To better understand the extent of mutagenesis in human bladder, the present study sequenced samples from normal urothelium.

Contributions

This chapter is based on the manuscript “Extensive heterogeneity in somatic mutation and selection in the human bladder” by Andrew R. J. Lawson, Federico Abascal, Tim H. H. Coorens, Yvette Hooks, Laura O’Neill, Calli Latimer, Keiran Raine, Mathijs A. Sanders,

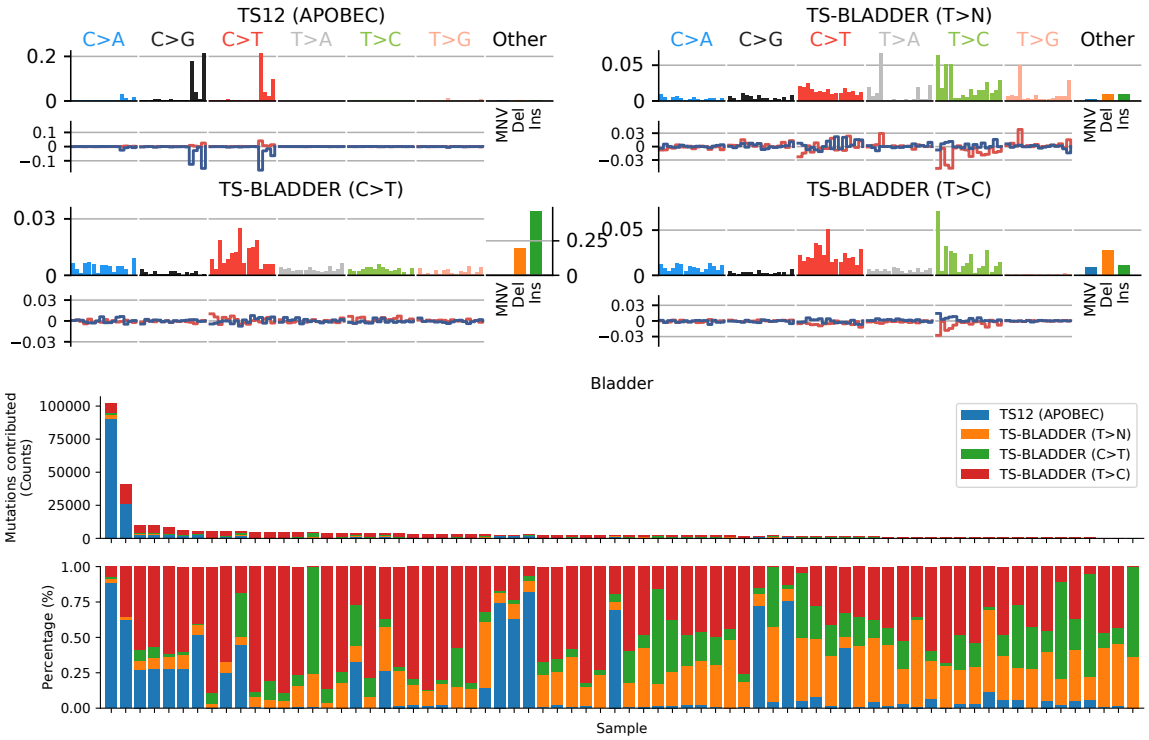


Fig. 3.31 **The tensor signatures and exposures of bladder normal urothelia.** Upper panel: The representation of tensor signatures analogous to Fig. 3.4. Lower panel: Absolute and relative signature contributions, each bar represents a sample.

Anne Y. Warren, Krishnaa T. A. Mahbubani, Bethany Bareham, Timothy M. Butler, Luke M. R. Harvey, Alex Cagan, Andrew Menzies, Luiza Moore, Alexandra J. Colquhoun, William Turner, Benjamin Thomas, Vincent Gnanapragasam, Nicholas Williams, Doris M. Rassl, Harald Vöhringer, Sonia Zumalave, Jyoti Nangalia, Jose M. C. Tubio, Moritz Gerstung, Kourosh Saeb-Parsy, Michael R. Stratton, Peter J. Campbell, Thomas J. Mitchell, Inigo Martincorena. H.V. validated the mutational signature analysis and produced the figures presented in this section.

The mutational signature analysis of normal urothelial revealed four tensor signatures

The tensor decomposition of normal bladder urothelia revealed four mutational processes (Fig. D.14). These comprise replication-driven APOBEC mutagenesis (TS12), as well as three novel tensor signatures TS-BLADDER (T>N), TS-BLADDER (C>T) and TS-BLADDER (T>C), whose spectra are characterised by T>N mutations, C>T mutations and indels, and T>C mutations with a transcriptional strand bias, respectively (Fig. 3.31, D.15). The great

majority of samples is characterised by high activities of TS-BLADDER (T>C), although the samples with highest mutation burden show substantial contributions from TS12.

3.3.5 The mutational processes of normal and tumorous cells

Cancer arises through the accumulation of somatic mutations throughout an individual's lifetime. For this reason, cells may acquire substantial amounts of mutations, often long before they become cancerous. Since the characterisation of mutational processes in normal tissues is likely to be important to better understand the development of cancer, the present study collected 1,511 normal and 664 tumorous samples from 981 individuals, comprising 11,226,385 SNVs and 18,895 MNVs.

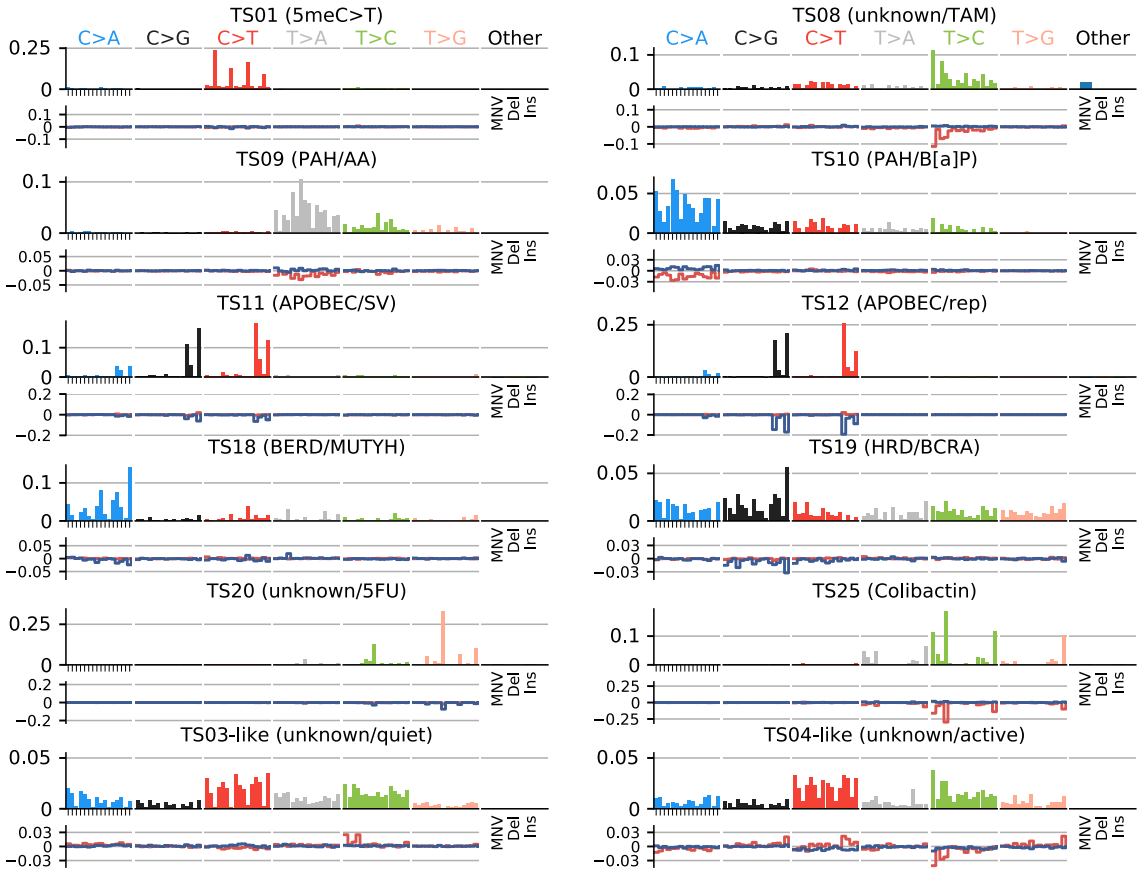


Fig. 3.32 **The tensor signatures of normal cells.** The representation of tensor signatures analogous to Fig. 3.4.

Contributions

This section contains preliminary results of an collaboration with Hyunchul Jung and contains the results of a manuscript in preparation. H.V. conducted all bioinformatic analyses and produced the figures presented in this section.

The TensorSignatures analysis of normal and tumorous cells revealed 12 tensor signatures

Applying TensorSignatures to the dataset of normal and tumorous cells uncovered twelve tensor signatures (Fig. D.16), ten of which closely resemble previously described tensor signatures. These mutational processes include 5-meC deamination (TS01), transcription-coupled damage (TS08), signatures indicative of exposure to aflatoxin and tobacco smoke (TS09 and TS10), mutagenesis due to SV and replication driven APOBEC mutagenesis (TS11 and TS12), signatures due to defects in BER and homologous recombination repair (TS18 and TS19), a cancer treatment associated signature (TS20), as well as a mutational pattern that strongly resembles the spectrum of the colibactin signature (TS25, Fig. 3.32). In addition, the analysis revealed two novel signatures TS03-like and TS04-like, whose base substitution spectra and epigenetic activation patterns remind of TS03 and TS04 (Fig. 3.32, D.17).

Distinct mutational signatures in cancer samples

The analysis of mutation counts revealed that tumours harbour 5 $\times$ greater mutational loads in comparison to normals (Median 1,029 vs. 5,789). However, although most normal tissue types exhibit mild mutational loads (median ranges from 574 in breast to 2,415 in colon), lung normal samples showed surprisingly high mutational burdens with a median of 102,306. In comparison, median mutation counts in tumours ranged from 3,035 in prostate to 40,844 in squamous lung cancers (Fig. D.18).

Next, I aggregated the exposures of identified tensor signatures to understand the contribution of different mutational processes across tissues (Fig. 3.33). The mutational spectrum of normal breast cells is characterised by spontaneous demethylation of 5-meC (25 %, Median: 129), as well as TS03 and TS04, which contribute 25 % (Median: 107) and 21 % (112) of mutations, respectively. Breast tumours, on the other hand, are dominated by mutagenesis due to APOBEC and failure of homologous recombination, contributing 29 % (293) and 22 % (641), respectively (Fig. 3.33, first row). Normal colon samples were dominated by TS01 contributing 48 % (Median: 1,118) of mutations, while the tumorous counterpart (colorectal adeno-carcinoma) was seemingly affected by defects in the BER pathway, as indicated by

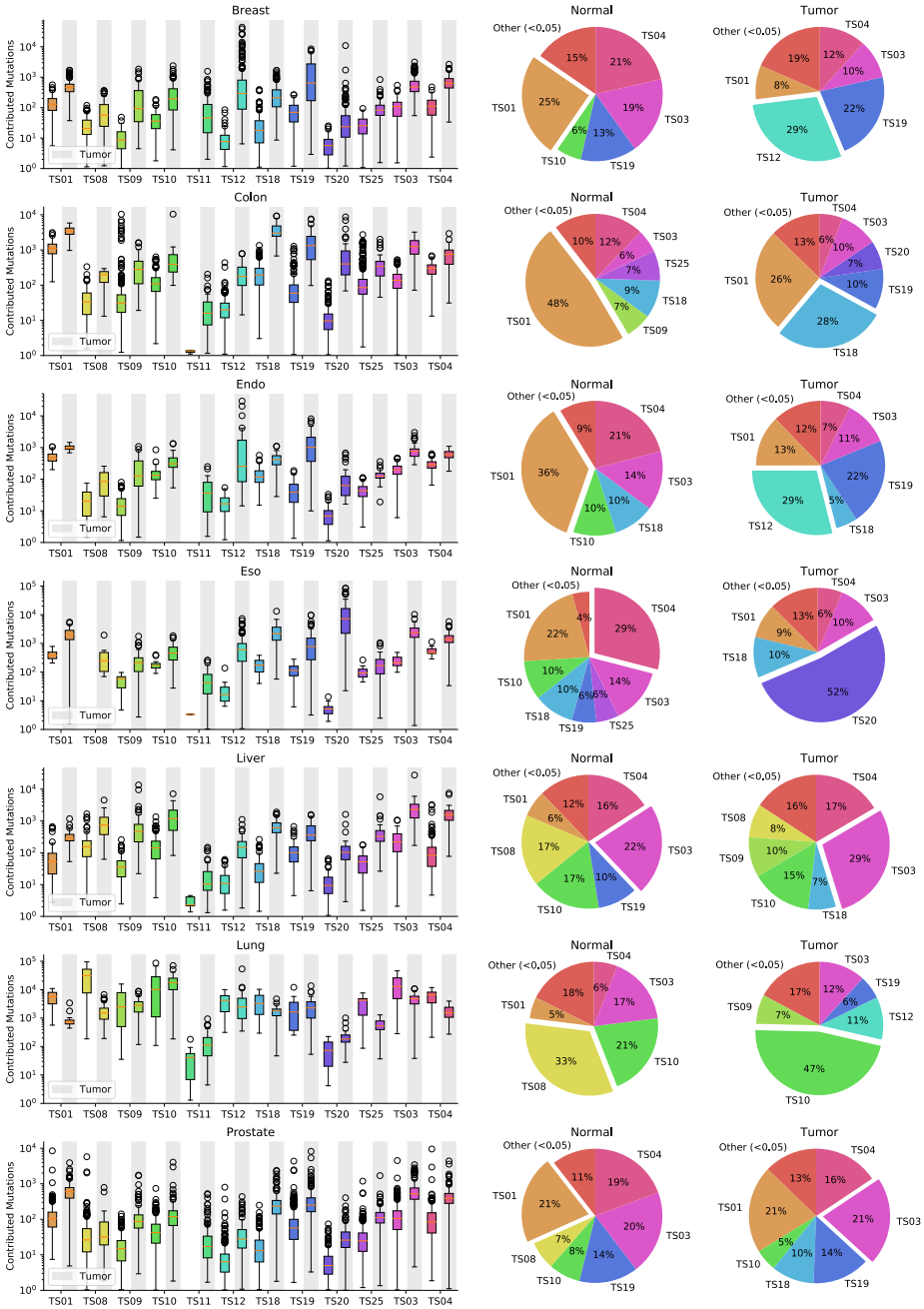


Fig. 3.33 **Aggregated exposures of normal and tumour samples.** Left column: The distribution of mutation counts of each mutational process in normals and tumours (breast carcinoma, colorectal adenocarcinoma, endometrial adenocarcinoma, oesophageal adenocarcinoma, hepatocellular carcinoma, squamous cell carcinoma and prostate adenocarcinoma) of each tissue. Right column: Overall contribution of each mutational process in normals and tumours of each tissue.

the large fraction of TS18 mutations (2,975, Fig. 3.33, second row). The composition of the mutational spectrum in endometrial samples is similar to breast, with TS01, TS03 and TS04 making up the majority of mutations in normals, and TS12 and TS19 in tumours, respectively (Fig. 3.33, third row). Oesophageal normals display the activity of many mutational processes, with TS04 being the most prominent, making up to 29 % (481) of mutations. In contrast, the mutational landscape in oesophageal adenocarcinoma is clearly dominated by TS20, which is responsible for more than half of the mutations (7,126), even though TS20 was hardly detectable in corresponding normals (Median: 5, Fig. 3.33, fourth row). Liver samples depicted a heterogenous mutational landscape, with substantial contributions of at least six mutational processes, and TS03 being the main contributor in normals and tumours (Median: 219 vs. 2,334, Fig. 3.33, fifth row). Normal lung samples showed high activity of the transcription coupled DNA damage associated process TS08, which dominated the mutational spectrum by contributing 33 % of mutations (Median: 22,860). Corresponding squamous lung cancers, however, predominantly exhibit TS10 mutations (47 %, Median: 18,555, Fig. 3.33, sixth row). Prostate normals and tumours were characterised by TS01 and TS03 activity, each contributing 21 % (Median: 105 and 522), respectively.

3.4 Summary

I presented TensorSignatures, a novel framework for learning mutational signatures jointly from their mutation spectra and genomic properties to better understand the underlying mutational processes. I illustrated the capabilities of this algorithm by presenting a set of 20 mutational signatures extracted from 2,778 cancer genomes of the PCAWG consortium, and validated my analysis on additional 3,824 metastatic samples from the HMF cohort. The number of signatures was deliberately kept low for the signatures to be interpretable. The analysis demonstrated that the majority of mutational signatures comprised different variant types, and that no single mutational signature acted uniformly along the genome. Measuring how mutational spectra are influenced by their associated genomic features sheds light on the mechanisms underlying mutagenesis. A joint inference also helps to dissect mutational processes in situations where mutation spectra are very similar, such that genomic associations cannot be unambiguously attributed based on the mutation spectrum alone.

1. Studying the resulting signatures revealed that the SNV spectra of TS05 and TS06 show high similarity to signatures SBS7a and SBS7b of the COSMIC catalogue of mutational signatures. Due to the high similarity of the mutational spectra, it is difficult to unambiguously attribute individual mutations to these signatures and measure their genomic activity and transcriptional strand biases based on the mutation spectra alone.

TensorSignature analysis reveals that the two processes are strongly differing with respect to their epigenetic context and transcriptional strand bias pointing towards differentially active GG-NER to be the underlying cause of the regional signature, which is confirmed by analysing cSCCs from GG-NER deficient XPC patients.

2. A similar change of the mutation spectrum was observed in Liver-HCC and other cancer types, reflected by the diverging activity of TS07 and TS08. The activity of TS08 is most prominent in highly transcribed genes, indicative of transcription-associated mutagenesis (Haradhvala et al., 2016; Imielinski et al., 2017). TensorSignatures unifies the overarching mutational spectrum of this process and sheds light on its genomic determinants. Furthermore, its ability to detect mutational signatures in specific genomic regions also increases the sensitivity to detect signature activity, which may only contribute low levels of mutation at a genome wide scale. Here, we find TS08 also in Bladder-TCC, ColoRectal-AdenoCa, Lung-AdenoCa, Prostate-AdenoCa and Stomach-AdenoCa in addition to Billiary-AdenoCa, Head-SCC, and Liver-HCC, where it has been previously found (Alexandrov et al., 2020).

3. TensorSignatures' capability to detect signatures with a confined regional context was also highlighted by detecting a highly localised signature associated with AID, TS13, which specifically manifests in and around transcription start sites in lymphoid neoplasms (Kasar et al., 2015). This signature has a base substitution spectrum similar to TS14 (SBS9), which does not display the tight localisation to TSS and is found in a range of cancer types, likely reflecting Polη-driven TLS during replication.

4. Inclusion of other mutation types led to the discovery of two APOBEC-associated signatures representative for mutagenesis during replication and at DSBRs, which differ with regard to their replicational strand bias and clustering propensity. Specifically, APOBEC-mediated mutagenesis at SVs lacks any preference for leading or lagging strand and is up to 80% clustered, suggesting that the formation of single stranded DNA during DSBR may trigger APOBEC activity. While an association of rearrangement events was reported earlier (Nik-Zainal et al., 2012), our study adds that DSBR- and replication-driven APOBEC mutations can be discerned by replication strand bias, clustering rate and size of clusters, indicating differential processivity of these two processes enabling different rates of mutation.

Validation of TensorSignature's predictions was achieved by applying the algorithm to a second cohort of whole genomes from the HMF. Reassuringly, the algorithm confirmed spectra and genomic properties from our primary analysis, thereby demonstrating the robustness of this approach, and the value of assessing additional genomic dimensions and

other mutation types to gain more in depth insights to mutagenesis. Moreover, I applied the software in various collaboration projects to smaller datasets, thus demonstrating the value of assessing the genomic properties of mutational processes.

Chapter 4

Discussion

The accumulation of genomic mutations eventually leads to the formation of cancer, a disease which is often associated with terminal outcomes for affected individuals. For this reason, many efforts have been undertaken to characterise mutational processes in terms of their genomic imprints. A particularly successful approach is matrix-based mutational signature analysis, which identifies prototypical mutation patterns by applying non-negative matrix factorisation to catalogues of single nucleotide variants and other mutation types. So far, the methodology has provided more than 50 different single base substitution patterns, indicative of a range of endogenous mutational processes, as well as genetically acquired hypermutation and exogenous mutagen exposures. However, mutagenesis is a multifaceted event that is affected by the genomic organisation of DNA and cellular processes such as transcription, replication, and DNA repair processes. Moreover, since many mutational processes also generate characteristic multi nucleotide variants, insertion and deletions, and structural variants, it appears valuable to jointly deconvolve broader mutational catalogues to better understand the complex nature of mutagenesis.

4.1 Summary of the main findings

In this thesis, I presented TensorSignatures, an algorithm to learn mutational signatures jointly across all variant categories and genomic context. I described how this method incorporates various mutation types and employs five different genomic annotations – transcription and replication strand orientation, nucleosomal occupancy, epigenetic states as well as local hypermutation – to create a seven-dimensional count tensor, enabling the algorithm to characterise mutational signatures with respect to the aforementioned phenomena. TensorSignatures is implemented using the powerful TensorFlow backend, benefits from GPU acceleration, and is available as Python package as well as a web application.

By applying TensorSignatures to 2,778 primary and 3,824 metastatic cancer genomes of the Pan Cancer Analysis of Whole Genomes consortium and the Hartwig Medical Foundation cohort, I found that practically all signatures operate dynamically in response to various genomic and epigenomic states. My analysis pinned differential spectra of UV mutagenesis found in active and inactive chromatin to global genome nucleotide excision repair. TensorSignatures accurately characterised transcription-associated mutagenesis, which is detected in 7 different cancer types. My analysis also unmasked replication- and double strand break repair-driven APOBEC mutagenesis, which manifests with differential numbers and length of mutation clusters indicating a differential processivity of the two triggers. As a fourth example, TensorSignatures detected a signature of somatic hypermutation generating highly clustered variants around the transcription start sites of active genes in lymphoid leukaemia, distinct from a more general and less clustered signature of Polη-driven translesion synthesis found in a broad range of cancer types. Finally, I demonstrate TensorSignatures' utility by applying it to multiple datasets in various collaboration projects.

Taken together, TensorSignatures adds great detail and refines mutational signature analysis by jointly learning mutation patterns and their genomic determinants. This sheds light on the manifold influences that underlie mutagenesis and helps to pinpoint mutagenic influences which cannot easily be distinguished based on the mutation spectra alone. As mutational signature analysis is an essential element of the cancer genome analysis toolkit, TensorSignatures may help make the growing catalogues of mutational signatures more insightful by highlighting mutagenic mechanisms, or hypotheses thereof, to be investigated in greater depth.

4.2 Conclusions

This analysis maps out the regional activity of mutational processes across the genome and pinpoints their various genomic determinants, thereby enabling to discover mutational processes not only in terms of their single base substitution spectra, but also with respect to their multi-facetted properties. This revealed that the appearance of a mutational signature is often driven by a combination of genomic determinants, other mutation types and intrinsic properties of the mutational process, e.g.

- epigenetic activity and transcriptional strand bias (TS05/TS06 and TS07/TS08),
- epigenetic activity and propensity to cluster (TS13/TS14),
- association with other mutation types, propensity to cluster and replicational strand bias (TS11/TS12).

Thus, providing a method with the capability to factor in all of these variables adds tremendous value to the field of mutational signature analysis. Particularly, it enables to

- generate a more informed hypothesis about the attribution of a mutational signature to its underlying origin.
- unravel interesting phenomena associated with certain mutational processes, helping to guide further downstream analysis.
- simplify the analysis by providing a general framework for the methodology, which in addition is easily extensible to account for different questions.

4.3 Limitations of the analysis and potential improvements

TensorSignatures promises to provide a robust framework for mutational signature analysis that enables to characterise mutational processes with respect to different genomic factors and the full mutational imprint. Despite the tool being fully tested and usable, there are several issues which I could not, or only partially address during the time of my PhD.

Algorithmic improvements

One drawback of the tensor representation is the memory footprint which scales multiplicatively with each genomic dimension. To illustrate this, consider the mutation count tensor presented in this thesis, which features 3 transcriptional, 3 replicational, 16 epigenetic, 4 nucleosomal, and 2 clustering states, as well as 96 mutation types and n samples, giving rise to a multidimensional array with $3 \times 3 \times 16 \times 4 \times 2 \times 96 \times n$ entries. This results in a $\sim$2.6 GB large mutation count tensor for the PCAWG dataset ($n = 2778$), and raises the question whether the algorithm is capable to handle substantially more samples and/or additional genomic covariates. Possible solutions to this problem include distributing the workload to multiple devices, training the model in batches, or marginalising the mutation count tensor iteratively.

The first option implies to distribute parts of the data to distinct GPUs, which then independently compute a single epoch including a forward and backward pass, after which gradients are collected and applied to the parameters of the model. In this way, the memory load is decreased on individual devices, thus enabling to train the model, even when the number genomes is large. However, this solution does not really solve the problem, because at some stage there will be surely enough data to overwhelm even the most sophisticated hardware setup. Moreover, this setup has other drawbacks, for example, collecting gradients

after each epoch requires to synchronise devices, which often results in poorer performance, especially when workloads are distributed unequally across GPUs.

The second approach may be implemented with batch- or mini-batch gradient descent. The former loads smaller batches of data to memory, calculates their log-likelihood and accumulates gradients, and updates the model after all training examples have been evaluated. While this is indeed a feasible solution, it makes the inference significantly slower, because loading batches into (GPU-) memory involves slow I/O-operations. Similarly, mini-batch gradient descent splits the dataset into small batches that are used to calculate the model log-likelihood and gradients, but directly updates model coefficients. In theory, the gradients computed from the batch represent noisy estimates of the true gradient, which should enable the algorithm to converge as if the gradient was computed on the complete dataset, thus allowing the model to converge faster.

The final proposal makes use of TensorSignatures' parametrisation, which models the effect of each genomic factor independently. This should allow to marginalise each genomic factor of the count tensor at a time, such that in addition to transcriptional and replicational states, as well as the dimensions for mutation types and samples, only one additional genomic dimension remains (e.g. $\sum_{\text{nuc, clu}} \mathbf{C}^{\text{SNV}} = \mathbf{C}^{\text{SNV}}_{\text{epi}} \in \mathbb{N}_0^{3\times3\times16\times p\times n}$, $\sum_{\text{epi, clu}} \mathbf{C}^{\text{SNV}} = \mathbf{C}^{\text{SNV}}_{\text{nuc}} \in \mathbb{N}_0^{3\times3\times4\times p\times n}$, and $\sum_{\text{epi, nuc}} \mathbf{C}^{\text{SNV}} = \mathbf{C}^{\text{SNV}}_{\text{clu}} \in \mathbb{N}_0^{3\times3\times2\times p\times n}$). Each marginalised count tensor has a significantly smaller memory footprint in comparison to the full tensor, enabling to store them all in memory, and train the model by iterating over each marginalised count tensor while accumulating gradients, and applying them after each such iteration.

Due to decreasing cost of sequencing and the thrive of personalised medicine, it is not too unlikely that whole genome sequencing becomes part of the standard clinical routine. While these developments bring the great opportunity to learn novel mutational processes from vast amounts of data, they will also represent a computational challenge to current NMF algorithms. Particularly, TensorSignatures tensor representation of mutation count data may soon exceed the memory boundaries of currently available hardware. To avoid these limitations, it is necessary to develop novel optimisation techniques for NMF, eventually based on batch gradient descent, or in the case of TensorSignatures, by marginalising the count tensor.

The TensorSignatures API

Regarding the current implementation of the TensorSignature software, I find two issues worth following up. First, the TensorSignature software requires to run a R pipeline to compute the SNV count tensor and the other mutation type matrix prior to the actual analysis, thus making the usage of the software difficult, as the installation of R and Bioconductor

packages depends on a plethora of dependencies. This could be resolved by performing these steps in Python, or providing a Docker image dedicated to this task (which is the current solution). Second, TensorSignatures was written using Tensorflow 1.x, which has received a major update in the meantime, suggesting to rewrite the software to make use of the Tensorflow 2.0 API. Also, I optimised the command line interface, as well as the API of the tensorsignatures Python package for the usage on high performance computing clusters, which question many of the design choices of the API when the software is run in different environments. With regards to theses issues, more development of the software needs to be done.

TensorSignatures Online

Cloud computing technologies offer interesting possibilities regarding the deployment of research pipelines, and thus many ideas come to mind to further improve the features of TensorSignatures' web application. At the time of writing, TensorSignaturesOnline allows to fit the exposures of a set of reference signatures to user samples. An evident extension is to provide the functionality to perform denovo signature fits, thus allowing users to extract the mutational processes present in their datasets.

Other ways to improve the web application involve its implementation and user experience. While the current version of TensorSignaturesOnline uses the Python web framework Flask to render the views of the page, it may be beneficial to provide a decoupled REST-API service, that is accessible without having to navigate to the webpage. This would enable to implement a Python independent front-end, for example, on basis of the Javascript framework React, which allows to build more elaborate user interfaces, thus making the usage of the application more convenient and intuitive.

Cancer specific annotations

Currently, TensorSignatures uses consensus annotations to classify single base substitutions with respect to replicational orientation, epigenetic and nucleosomal states. This simplifies data preprocessing steps, i.e. a single normalisation constant may be employed to account for different nucleotide compositions across genomic states, and provides conservative estimates for inferred tensor factors, because assignment of single base substitutions is based on conserved genomic regions. However, the usage of consensus states may also preclude signal due to the assignment of mutations to the NA state when cell-type specific genomic states fail to reach the threshold for the consensus.

While currently available data does not allow to obtain a comprehensive set aforementioned annotations matched to the cell of origin for every PCAWG cancer type, it is possible to partially match epigenetic ChromHMM annotations, which is the genomic variable having arguably the greatest impact on mutagenesis. Although I found that the analysis on partially matched annotations yields concordant results (Fig. B.3), it may be desirable to use fully matched annotations in future.

Additional genomic states

TensorSignatures' extensibility enables to assess mutational processes with respect to any genomic annotation. Some interesting factors could be easily gathered from standard genomic data such as chromosomes, introns, or exons; but certainly from more specific annotations as well, for example, chromosomal contact sites captured by Hi-C data. To illustrate this, consider the following two examples.

The genome consists of several higher order DNA domains, occupying different territories within the nucleus, either in the periphery close to the nuclear envelope, or central at its core. It has been reported that the availability of DNA repair mechanisms differ between these higher order domains. Particularly, nuclear-lamina-associated regions are more likely to use error prone alternative non-homologous end joining, and are less efficient in recruiting the proteins of the NER-pathway to lesions, as only XPC allocates to the nuclear periphery but not XPA (Sec. 1.4.2). For this reason, different mutational signatures were observed across high order DNA domains (Smith et al., 2017). However, a more thorough analysis might be easy to achieve using TensorSignatures, for example, by utilising genomic annotations for Chip-Seq data gathered from lamin proteins to partition single base substitutions by nuclear localisation.

While TensorSignatures may resolve mutagenesis in large genomic regions, it may also be used to detect mutational processes at smaller scales. Another interesting property of DNA is its propensity to form secondary structures at certain motifs. In particular, guanine rich sequences have the propensity to form so-called G-quadruplexs (G4s), which are often found in telomeric regions but also at gene regulatory loci. G4 may obstruct the movement of DNA polymerases, thus increasing the risk of DNA breakage, and have been shown to associate with copy number alterations. Regions with the propensity to generate G4s could be introduced as an additional genomic dimension to the mutation count tensor, enabling to assess the mutational processes acting at these genomic regions. However, the incorporation of additional genomic states should be accompanied by larger catalogues of cancer genomes to preserve the statistical power of the analysis.

Tensor for other mutation types

Another possibility to further improve the tensor factorisation approach is to partition other mutation types with respect to various genomic annotations. The difficulty herein is that annotating a subset of mutation types such as structural variants, larger MNVs and indels with respect to various genomic states is not straightforward. To understand this, consider that sites of structural variation usually affect both DNA strands, thus making it impossible to assign them with regards to transcription or replicational directionality. On the other hand, larger MNVs and indels may overlap with two adjacent genomic states, thereby making unambiguous state assignments difficult. These issues may be partly resolved by finding more elaborate mutation type classifications.

Partitioning other mutation types by genomic factors, especially indels, could greatly help to make sense out of the large number of mutational signatures that have been attributed to mismatch repair deficiency (MMRD, Sec. 1.6.2). Cancers harbouring defects in this DNA repair pathway are associated with the formation of indels, especially at sites of repetitive DNA sequences, and are therefore called microsatellite instable. Moreover, it has been reported that MMR is underlying differential mutation rates across early and late replicating regions. Stratifying indels by epigenetic factors could therefore help to disentangle the association of certain MMRD signatures and link them to genomic factors.

Metadata

Another exciting avenue to drive forward the field of mutational signature analysis is to extend the method to incorporate patient meta data such as age, treatment history, smoking status etc. Such data could be incorporated in secondary matrices, much like the incorporation of other mutation types in TensorSignatures, thus helping to guide signature inference at the stage of training. Concrete parameterisations of such models need yet to be developed, and are likely to depend on the data type added, but could help to link the origin of unknown or novel signatures to their respective origin.

4.4 Outlook and future research

Although the origins of mutational signature analysis lie in the decomposition of (primary) cancer genomes, it is evident that cancer development is a continuous process that eventually starts in a normal cell. An interesting avenue of research is therefore to assess the mutational composition of normal cells at consecutive stages during cancer development to learn more about the different phases of carcinogenesis. Indeed, various recent studies started to delineate

the mutational composition of normal cells in different tissues including liver, colorectal, bronchial, bladder and endometrial epithelia (Brunner et al., 2019; Lee-Six et al., 2019; Moore et al., 2020; Yoshida et al., 2020), but further research is required to obtain a more comprehensive view on the activity of mutational processes in normals, and at certain stages of tumor development. Likewise, it may be beneficial to sequence more metastatic tumours, albeit the HMF study provided a quite comprehensive catalogue of approximately 4,000 samples (Priestley et al., 2018).

The human body consists of approximately 40 trillion cells and more than 100 different cell types. Given this large diversity, it is not surprising that most tissues as well as malignant tumours contain a plethora of different cell types. Nonetheless, conventional sequencing experiments typically explore cells in bulk, for which reason the results of such experiments reflect entire cell populations rather than a single cell. The reason for the lack of resolution is that the amount of DNA that can be extracted from a single cell is often not sufficient to enable genome-scale analysis, even though such information could help delineate the clonal aetiology of cell sub populations in cancers. Recent advances in single cell genomics could help to unravel the signals of such samples, thereby delivering cleaner signals that may help, for example, to associate certain mutational signatures with distinct cell types.

Despite the temporal and spatial characterisation of mutational processes in tumours, integration of other ”omics” data from the epigenome, transcriptome, proteome, metabolome and microbiome may further help to make mutational signature analysis more powerful. While TensorSignatures took first steps to integrate data from the epigenome and transcriptome, my consensus approach is likely to be overwhelmingly crude in comparison to the resolution that may be gathered from profiling tumours with respect to these modalities. In addition, proteomics data, which is used to quantify peptide abundance, interaction and modification, may help to uncover mechanistic details of the action of tumour surpressors and oncogenes by unravelling their abundance and state as protein. Metabolomics quantifies multiple small molecules, such as amino acids, fatty acids and other products of the cellular metabolism, and may therefore help to elucidate the genomic consequences of exogenous mutagens, as well as the effects of an altered metabolism in deregulated cancerous cells. Finally, integration of the microbiome may help to unravel the interactions between (epithelial) cancers and the surrounding microenvironment.

Many of the mutational signatures present in the COSMIC catalogue lack an attribution to a concrete mutational process. To learn more about the aetiology of unknown spectra, multiple studies simulated carcinogenesis in various model organisms on the bench, including genetically modified *C. elegans* (Meier et al., 2018; Volkova et al., 2019), CRISPR-Cas9 modified iPSC and cancer cell lines (Petljak et al., 2019; Zou et al., 2018), as well as human

organoids (Drost et al., 2017). Each of these approaches have their pros and cons: studies conducted in *C. elegans*, for example, enable to perform large scale screens, i.e. it is possible to expose genetically modified *C. elegans* with defects in tumour suppressive or oncogenic genes to a wide range of mutagenic substances, which helps to identify the mutational spectra of interactions, but obviously the model system is evolutionary distant from humans. In contrast, CRISPR-Cas9 modified cell lines as well as organoids are more likely to reflect mutagenesis in human cells, but require very laborious and expensive experiments. Moving forward, further experimental work is required to complement and validate the advances and findings in the field of mutational signature analysis, ideally in experimental setups that are able to reflect the genomic complexity of the human genome.

Appendix A

TensorSignatures Manual

TensorSignatures is a tensor factorisation framework for mutational signature analysis, which in contrast to other methods, deciphers mutational processes not only in terms of mutational spectra, but also assess their properties with respect to various genomic variables, allows the inclusion of different mutation types and integrates a robust noise model to perform the inference.

TensorSignatures is a young project and breaking changes are to be expected. We keep a changelog and it will have possible breakage clearly documented.

A.1 Installing TensorSignatures

TensorSignatures makes use of the TensorFlow 1.5.x framework requiring the user to install a separate package to enable GPU support, i.e. `tensorflow-gpu` instead of `tensorflow`. We highly recommend to install TensorSignatures into an environment with `tensorflow-gpu`, as the tensor computations greatly benefit from GPU-acceleration.

A.1.1 Installation via GitHub

To obtain the most recent version of TensorSignatures, we recommend to download the repository directly from GitHub and to install the package into a virtual environment. To get started, clone the repository by executing the following commands in your terminal

```
$ git clone https://github.com/gerstung-lab/tensorsignatures.git && \
cd tensorsignatures
```

Then, create a new virtual environment and install all dependencies. If you have access to a GPU with cuda support use `requirements-gpu.txt` instead of `requirements.txt`

```
$ python -m venv env
$ source env/bin/activate
$ pip install --upgrade pip setuptools wheel && \
pip install -r requirements.txt
```

Finally, install TensorSignatures.

```
$ python setup.py install
```

A.1.2 Installation via Pypi

To install `tensorsignatures` via Pypi simply type

```
$ pip install tensorsignatures
```

into your shell.

A.1.3 Installation via Docker

To run TensorSignatures within a docker environment clone the the repository

```
$ git clone https://github.com/gerstung-lab/tensorsignatures.git
$ cd tensorsignatures
```

and then start the image using `docker-compose`.

```
$ docker-compose up --build
```

This spins up a jupyter server including notebooks with tutorials on `http://localhost:8889`.

A.2 Quick Start

Running TensorSignatures involves three steps: preparing the input data, i.e. creating the mutation count tensor as well as the mutation count matrix, computing a trinucleotide normalisation to account for differences in the nucleotide composition of different genomic regions, and running TensorSignatures.

A.2.1 Step 1: Data preparation

Preparing the input data for TensorSignatures involves creating the single base substitution count tensor and the other mutation type count matrix with multinucleotide variants, deletions and insertions (currently we do not provide a automated way of generating a structural variant table yet). Despite the fact that TensorSignatures is written in Python, this part of the pipeline runs in `R` and depends on the `Bioconductor` packages `VariantAnnotation` and `rhdf5`.

Preparing input data using docker

We provide a docker image that contains all `R` and `bioconductor` dependencies to create the variant tensor and the other mutation type matrix. To use it, pull the image from docker. Note that the image is approximately 5 GB large.

```
$ docker pull sagar87/tensorsignatures-data:latest
```

To use the image switch into the folder containing your VCF data. Then run image using the following command and supply the VCF files as well as the name of the `hdf5` output file (must be the last argument) as arguments.

```
$ docker run -v $PWD:/usr/src/app/mount sagar87/tensorsignatures-data \
<vcf1.vcf> <vcf2.vcf> ... <vcfn.vcf> <output.h5>
```

Then continue with Step 2 (Sec. A.2.2).

Preparing the input data using a custom installation

Make sure you have `R3.4.x` (!) and the packages `VariantAnnotation` and `rhdf5` installed. You can install them, if necessary, by executing

```
$ Rscript -e "source('https://bioconductor.org/biocLite.R'); \
biocLite('VariantAnnotation')"
```

and

```
$ Rscript -e "source('https://bioconductor.org/biocLite.R'); \
biocLite('rhdf5')"
```

from your command line. Then, download the following files and place them in the same directory:

- `Constants.RData` (contains `GRanges` objects that annotate transcription/replication orientation, nucleosomal and epigenetic states)
- `mutations.R` (all required functions to partiton SNVs, MNVs and Indels)
- `processVcf.R` (loads vcf files and creates the SNV count tensor, MNV and indel count matrix; eventually needs custom modification to make the script run on your vcfs.)
- `genome.zip`.

To obtain the SNV count tensor and the matrices containing other mutation types, execute `processVcf.R` and pass the VCF files you want to convert, as well as a name for an output `hdf5` file as command line arguments, e.g.

```
$ Rscript processVcf.R <vcf1.vcf> <vcf2.vcf> ... <vcfn.vcf>\
<output.h5>
```

In case of errors please check wether you have correctly specified paths in line 6-8. Also, take a look at the `readVcfSave` function and adjust it when it fails.

A.2.2 Step 2: Computing trinucleotide normalisation

TensorSignatures requires a trinucleotide normalisation constant to account for differences in the nucleotide composition of genomic states. To compute it, invoke the prep sub routine of TensorSignatures and pass the `hd5` file from Step 1 (Sec. A.2.1) as well as the path for the output file as positional arguments to the programme.

```
$ tensorsignatures prep <output.h5> <tsdata.h5>
```

A.2.3 Step 3: Run TensorSignatures

There are two ways to run TensorSignatures using either the `refit` option, which fits the exposures of a set of pre-defined signatures extracted from the PCAWG cohort to your dataset, or via the `train` subroutine, that performs a denovo extraction of tensor signatures. Refitting tensor signatures is computationally fast but does not allow to discover new signatures, while extracting new signatures from scratch is computationally intensive (GPU required) and requires ideally larger numbers of samples. For most use cases, with a small number of samples, we advice to use the refit option:

```
$ tensorsignatures --verbose refit tsData.h5 refit.pkl -n
```

To run a denovo extraction use

```
$ tensorsignatures --verbose train tsData.h5 denovo.pkl <rank> \
-k <size> -n -ep <epochs>
```

where `rank` specifies the decomposition rank, `size` controls the dispersion of the model, and `epochs` the number of desired epochs to fit the model. TensorSignatures outputs value of the objective function (log likelihood) that is minimised during training as well as the change of the objective during an epoch interval (`delta`). When deciding on the number of epochs to train the model ensure that it is sufficiently large such that the objective function converges, i.e. the `delta` value is close to, or fluctuates around zero. For more information on how to run TensorSignatures in a practical setting see Sec. (A.3.3). Running TensorSignatures will yield a `pickle` dump which can subsequently inspected using the `tensorsignatures` package (A.3.4).

A.3 Tutorials

TensorSignatures extracts mutational signatures and their genomic properties from a mutation count tensor that partitions single base substitutions with respect to a multitude of genomic states. Moreover, the algorithm links other variant types to these signatures by taking into account a secondary mutation matrix. In following tutorials, we want to convey an intuition for working with such highdimensional data, and explain the usage of the `tensorsignatures` API and command line interface (CLI).

A.3.1 Understanding the mutation count tensor

Let us start by creating a simulated mutation count tensor to better understand the structure of this data type. The `data` module of `tensorsignatures` provides a `TensorSignatureData` class which allows us to simulate such data. To do so, we import the package and create a `TensorSignatureData` instance.

```
> import tensorsignatures as ts
> data = ts.TensorSignatureData(
    seed=573, # set a seed for reproducibility
    rank=5, # number of signatures
    samples=100, # number of samples
```

```
    dimensions=[3, 5], # number of arbitrary genomic dimensions
    mutations=1000)
```

This command generates data from five signatures (`rank`) to simulate 100 genomes (`samples`) each with 1000 mutations (`mutations`). By passing the list [3, 5] to the `dimensions` argument, we create two additional genomic dimensions[1] with 3 and 5 states respectively. To obtain the SNV count tensor, we invoke the `snv` method of `data`, which returns the single base substitution count tensor.

```
> snv = data.snv()
```

Similarly, we can extract a simulated matrix of other mutation counts by invoking the `other()` method.

```
> other = data.other()
```

The mutation count tensor is a multidimensional array with a specific structure

The `snv` object is simply a 6-dimensional `numpy` array,

```
> snv.ndim
6
```

whose `shape` attribute is a tuple of integers indicating the size, i.e. the number of states, of the array in each dimension.

```
> snv.shape
(3, 3, 3, 5, 96, 100)
```

TensorSignatures expects the structure of the count tensor to follow a specific convention: the first and second dimension (`snv.shape[0]` and `snv.shape[1]`) split counts by transcription and replication strand, following dimensions partition single base substitution by genomic factors, and the penultimate (`snv.shape[-2]`) and last dimension (`snv.shape[-1]`) represent substitution types and samples respectively. Table A.1 summarises the structure of the count tensor.

[1] in addition to the dimensions specifying transcription and replication

Table A.1 The structure of the SNV count tensor.

Dimension	Size	Index	State / Variants
Transcription	3	0	Coding strand
		1	Template strand
		2	Unassigned
Replication	3	0	Leading strand
		1	Lagging strand
		2	Unassigned
First genomic dimension (eg. epigenetic states)	t+1	0	Unassigned
		1	State 1
		...	...
		t	State t
Last genomic dimension (eg. nucleosomal states)	r+1	0	Unassigned
		1	State 1
		...	
		r+1	State r
Single base substitution types	p=96	0	A[C>A]A
		1	A[C>A]C
		...	...
		p	T[T>C]T
Samples	n	0	Sample 1
		...	
		n	Sample n

Extracting the single base substutions from specific genomic states

We index the SNV tensor like any other `numpy` array. For example, to obtain variants from template and leading strands, and from the "unassigned" state of additional genomic dimensions, we simply index the tensor with `snv[0, 1, 0, 0, :, :]` which returns a two dimensional array with mutation types along the first axis and samples along the other.

```
> slice = snv[0, 1, 0, 0, :, :]
> slice.shape
(96, 100)
```

Note, that we can reconstruct the $p \times n$ mutation count matrix, which usually serves as an input for conventional mutational signature analysis, by summing over all dimensions

except the last two (representing single base substitution types and samples respectively). The following code illustrates this operation.

```
> collapsed = snv.sum(axis=(0, 1, 2, 3))
> collapsed.shape
(96, 100)
```

Another useful technique is to first index a specific state, and then to sum over all other dimensions to exclude. This allows us to extract the spectra from specific genomic states, for example, to extract all coding and template strand mutations from the tensor we would simply run

```
> coding = snv[0].sum(axis=(0, 1, 2, 4))
> template = snv[1].sum(axis=(0, 1, 2, 4))
```

of course this also works for any other dimension, for example, leading and lagging strand mutations maybe extracted as follows.

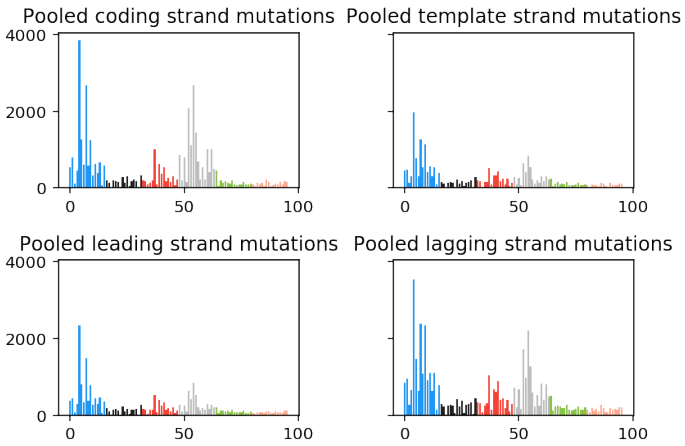
```
> leading = snv[:, 0].sum(axis=(0, 1, 2, 4))
> lagging = snv[:, 1].sum(axis=(0, 1, 2, 4))
```

To understand how they differ we may plot them,

```
> fig, axes = plt.subplots(2, 2, sharey=True)
> axes[0, 0].bar(np.arange(96), coding, color=ts.DARK_PALETTE)
> axes[0, 0].set_title('Pooled coding strand mutations')
> axes[0, 1].bar(np.arange(96), template, color=ts.DARK_PALETTE)
> axes[0, 1].set_title('Pooled template strand mutations')
> axes[1, 0].bar(np.arange(96), leading, color=ts.DARK_PALETTE)
> axes[1, 0].set_title('Pooled leading strand mutations')
> axes[1, 1].bar(np.arange(96), lagging, color=ts.DARK_PALETTE)
> axes[1, 1].set_title('Pooled lagging strand mutations')
> plt.tight_layout()
```

which reveals that some variant types, e.g. C>A (blue), C>T (red) and T>A (grey), seem to occur with different frequencies across transcription and replication states.

By indexing the SNV tensor appropriately, we can also recover mutational spectra from different state combinations, eg. `snv[0,:,2].sum(axis=(0,1))` would return a $p \times n$ matrix representing the coding strand mutations in state 2 of the first additional genomic dimension.



A.3.2 Understanding tensor factors

In the previous section, we created a simulated dataset using the `TensorSignaturesData` class, and investigated the data by plotting mutational spectra in various genomic contexts. While doing so, we discovered that some variant types occur with different frequencies in different genomic states, for example, frequencies of coding strand C>A, C>T and T>A variants seemed to be twice as large in comparison to corresponding numbers on template strand DNA. Strand asymmetries have been observed for several mutational processes and are often attributed to DNA repair mechanisms. Transcription coupled repair (TCR), for example, actively depletes mutations on template strand DNA in gene encoding regions.

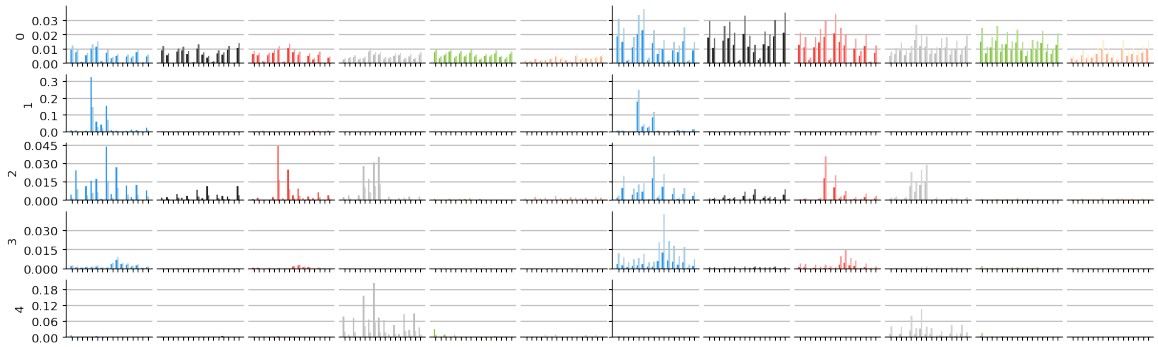
Transcriptional and replicational biases

TensorSignatures models variability in mutagenesis due to transcription and replication by

1. extracting separate single base substitution spectra for coding and template strand, and leading and lagging strand DNA
2. fitting a scalar for each signature in context of transcription and replication that quantifies the overall strand asymmetry of single base substitutions (bias matrix `b`)
3. fitting a scalar for each signature that is interpreted as the relative signature activity of signature in transcribed vs untranscribed regions, and early and late replicating regions (activity matrix `a`).

To understand this better, let us first plot the signatures that were used to simulate the counts in `data`.

```
> plt.figure(figsize=(16, 5))
> ts.plot_signatures(data.S.reshape(3, 3, -1, 96, data.rank))
```

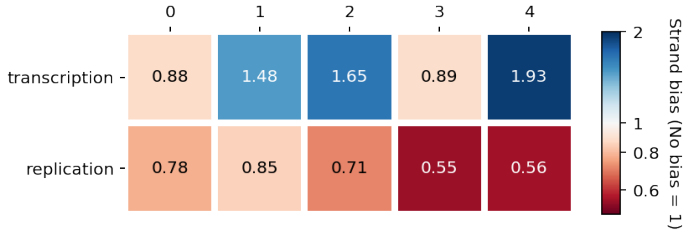


This reveals the SNV spectra (rows) in context of transcription and replication in the left and right column. Colors indicate the mutation type (blue C>A, black C>G, red C>T, grey T>A, green T>C and salmon T>G), while the shading indicates the mutation type probabilities for coding strand and leading strand DNA (dark), and for template and lagging strand DNA (light), respectively. Notice, for example, how in the fourth signature (last row), the amplitude of dark and light grey bars differ, indicating that this mutational process is more likely to produce T>A mutations on coding and leading strand DNA respectively.

TensorSignatures models the propensity of a mutational process to generate strand specific mutations by scaling the SNV spectra for coding and template, and leading and lagging strand with a multiplicative scalar variable. To visualise the strand biases for our simulated dataset, we pass the strand biases, accessible via the `b` attribute of our `data` object, to the `ts.heatmap` function.

```
> plt.figure(figsize=(6,2))
> ts.heatmap(data.b,
           vmin=.5, vmax=2, # allows to specify the limits of the colorbar
           row_labels=['transcription', 'replication'],
           cbarlabel='Strand bias (No bias = 1)' # color bar label
          )
```

Rows of the heat map depict the context and columns signatures. Note the logarithmic scaling of the color bar, which indicates that a baseline value of 1 resembles a mutational process with no strand preference. Coefficients < 1 (red) indicate signature enrichment on template or lagging strand DNA, and conversely, values > 1 (blue), an asymmetry towards the on coding or leading strand.



Signature activities in specific genomic regions

The multidimensional representation of SNV count data allows TensorSignatures to quantify the propensity of mutational processes within confined genomic regions. These genomic contexts, thereafter also genomic states, may represent genomic features such as specific chromatin marks or nucleosome occupancy. To illustrate this, we depicted a genomic region in the Fig. A.1 together with arbitrary genomic states and respective mutations.

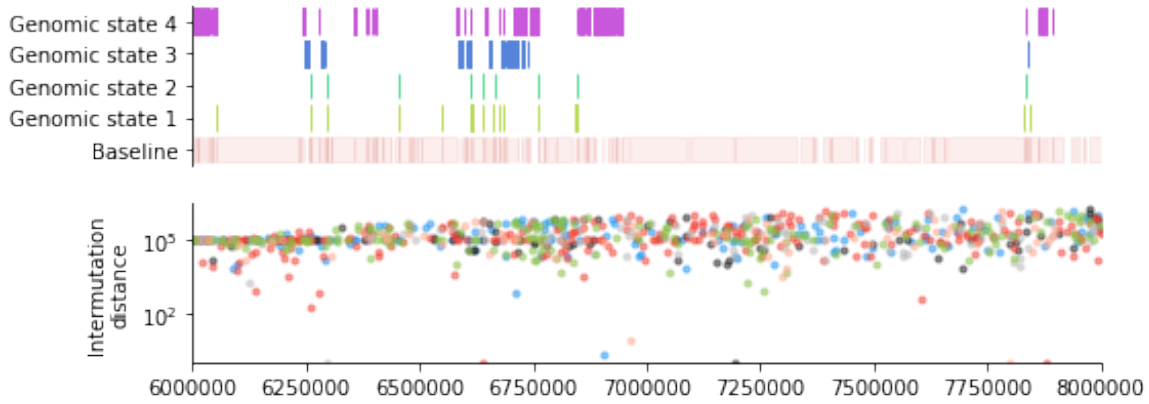
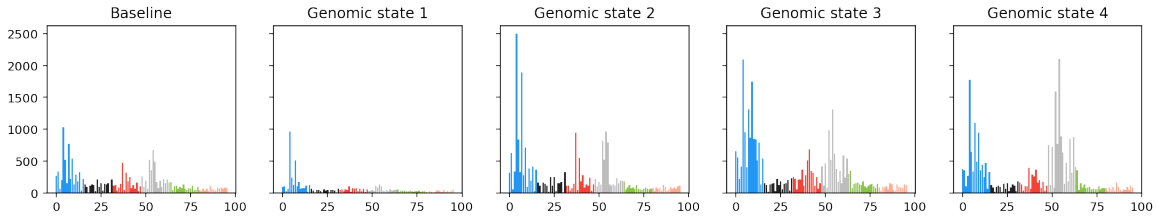


Fig. A.1 **The distribution of single base substitutions may vary due to differences in genome organisation and other factors.** The horizontal bar plot in the upper panel depicts genomic states, which represent confined genomic regions with certain features. The rainfall plot underneath shows the variant types at these genomic loci.

The rainfall plot representation may not always reveal changes in the mutational spectrum on first sight. However, the SNV count tensor contains the mutational spectra of each state combination. We can inspect them by indexing the respective state and summing over all remaining dimensions except the one for trinucleotides (Sec. A.3.1). To visualize, for example, pooled mutation spectra along the five states of the fourth dimension in our simulated dataset we would execute the following code.

```
> fig, ax = plt.subplots(1, 5, figsize=(16, 2.5), sharey=True)
> ax[0].bar(np.arange(96), snv[:, :, :, 0].sum(axis=(0, 1, 2, 4)), \
```

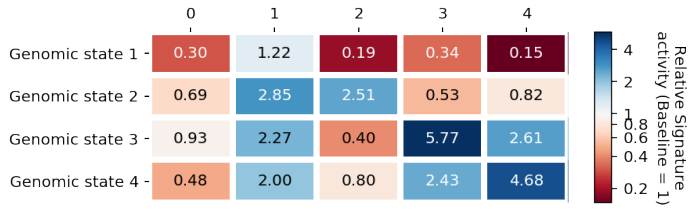
```
color=ts.DARK_PALETTE)
> ax[0].set_title('Baseline')
> ax[1].bar(np.arange(96), snv[:, :, :, 1].sum(axis=(0, 1, 2, 4)), \
color=ts.DARK_PALETTE)
> ax[1].set_title('Genomic state 1')
> ax[2].bar(np.arange(96), snv[:, :, :, 2].sum(axis=(0, 1, 2, 4)), \
color=ts.DARK_PALETTE)
> ax[2].set_title('Genomic state 2')
> ax[3].bar(np.arange(96), snv[:, :, :, 3].sum(axis=(0, 1, 2, 4)), \
color=ts.DARK_PALETTE)
> ax[3].set_title('Genomic state 3')
> ax[4].bar(np.arange(96), snv[:, :, :, 4].sum(axis=(0, 1, 2, 4)), \
color=ts.DARK_PALETTE)
> ax[4].set_title('Genomic state 4')
```



This plot nicely illustrates that different genomic states may have a variable exposure to different mutational signatures. For example, judging from the prevalence of C>A and T>A variants in genomic state 2 and 4, it appears likely that these states are dominated by signature 3 and 4 respectively. TensorSignatures models the activity of each signature by fitting a single coefficient for each signature and genomic state. To visualize the coefficients used to generate our simulated dataset we execute

```
> plt.figure(figsize=(6,2))
> ts.heatmap(data.k1,
           row_labels=['Genomic state 1', ..., 'Genomic state 4', ],
           col_labels=['{}'.format(i) for i in range(5)],
           cbarlabel='Relative Signature\nactivity (Baseline = 1)')
```

which confirms our suspicion about the elevated activities of signature 3 and 4 in genomic state 3 and 4 respectively. To interprete this correctly, keep in mind that usually majority of SNVs do not fall into specific genomic states and therefore end up in the baseline or



"unassigned" state (Tab. A.1), which is in TensorSignatures always 1, and to which all other coefficients are inferred relatively to. In other words, signature 3, for example, shows 5.77 times higher activities in genomic state 3 in comparison to the genomic baseline.

A.3.3 The TensorSignatures CLI

The TensorSignatures CLI comes with six subroutines,

- `boot`: computes bootstrap intervals for a TensorSignature initialisation,
- `data`: simulates mutation count data for a TensorSignature inference,
- `prep`: computes a normalisation constant and formats a count tensor (Sec. A.2.2),
- `refit`: refits the exposures to set of fixed tensor signatures (Sec. A.2.3),
- `train`: runs a denovo extraction of tensor signatures (Sec. A.2.3),
- `write`: creates a hdf5 file out of dumped tensor signatures pkls.

The goal of this tutorial is to illustrate how to run TensorSignatures in a practical setting. For this reason we will first simulate mutation count data using `tensorsignatures data`, and subsequently run `tensorsignatures train` to extract constituent signatures. In the next section we will then analyse the results of this experiment in jupyter with help of the tensorsignatures API (Sec. A.3.4).

Simulate data via the CLI

To create a reproducible (the first positional argument sets a seed: 573) synthetic dataset from 5 mutational signatures (second positional argument) with the CLI, we invoke the data subprogram

```
$ tensorsignatures data 573 5 data.h5 -s 100 -m 10000 -d 3 -d 5
```

which will simulate 100 samples (`-s 100`) each with 10,000 mutations (`-m 10000`), and two additional genomic dimensions with 3 and 5 states (`-d 3 -d 5`) respectively. The program writes a `hdf5` file `data.h5` to the current folder containing the datasets `SNV` and `OTHER` representing the SNV count tensor and all other variants respectively.

Running TensorSignatures using the command line interface

Since we know the number of signatures that made up the dataset we can run a TensorSignatures decomposition simply by executing

```
$ tensorsignatures --verbose train data.h5 my_first_run.pkl 5
```

which saves a pickle able binary file to the disk, which we can load into a interactive python session (eg. a Jupyter notebook) for further investigation (see Sec. A.3.4)

```
> init = ts.load_dump('my_first_run.pkl')
> init.S.shape
(3, 3, 3, 5, 96, 5, 1)
```

However, usually we do not know the number of active mutational processes a priori. For this reason, it is necessary to run the algorithm using different decomposition ranks, and to subsequently select the most appropriate model for the data. Moreover, we recommend to run several initialisations of the algorithm at each decomposition rank. This is necessary, because non-negative matrix factorisation produces stochastic solutions, i.e. each decomposition represents a local minimum of the objective function that is used to train the model. As a result, it is worthwhile to sample the solution space thoroughly, and to pick the solution which maximised the log-likelihood. Running TensorSignatures at different decomposition ranks while computing several initialisations is easy using the CLI. For example, to compute decompositions from rank 2 to 10 with 10 initialisation each, we would simply write a nested bash loop.

```
$ for rank in {2..10}; do
$   for init in {0..9}; do
$     tensorsignatures train data.h5 sol_${rank}_${init}.pkl ${rank} \
-i ${init} -j MyFirstExperiment;
$   done;
$ done;
```

Also note the additional arguments we pass here to the programme; the `-i` argument identifies each initialisation uniquely (mandatory), and the `-j` parameter allows us to name the experiment, which in this context denotes multiple TensorSignature decompositions across a range of ranks extracted using the same hyper parameters (number of epochs, dispersion, etc).

Summarising the result from many initialisations with `tensorsignatures write`

This command produces for each rank (2-10) ten initialisation and saves the results as pickleable binary files to the hard disk. Loading the 9 x 10 initialisations manually using `ts.load_dump` would be quite tedious and even impracticable in larger experiments. For this reason, we included the subprogram `tensorsignatures write`, which takes a `glob` filename pattern and an output filename as arguments to generate a `hdf5` file containing all initialisations.

```
$ tensorsignatures write "sol_*.pkl" results.h5
Processing 90 files.
```

A.3.4 The TensorSignatures API

The TensorSignatures API provides useful functions to analyse results from TensorSignature decompositions. Since running the tool usually involves creating several initialisations at different decomposition ranks (Sec. A.3.3), we provide three classes that abstract

- Experiments (`Experiment`), i.e. multiple initialisation at different decomposition ranks extracted using the same hyper parameters,
- Cluster (`Cluster`), i.e. multiple initialisations at a specific decomposition rank,
- Initialisations (`Initialization`): a single decomposition.

Importing data and performing model selection using the `Experiment` class

The `Experiment` class loads and clusters initialisations of each decomposition rank of a `hdf5` file written by `tensorsignatures write` (Sec. A.3.3)

```
> experiment = ts.Experiment("results.h5")
```

The `data` field of an `Experiment` instance returns a set of keys, which allow us to access the `Cluster` of the experiment (Sec. A.3.4). Keys follow the format to prefix the

decomposition rank with the name of the experiment, which we set earlier using the `-j` flag when we ran `tensorsignatures train` (Sec. A.3.3).

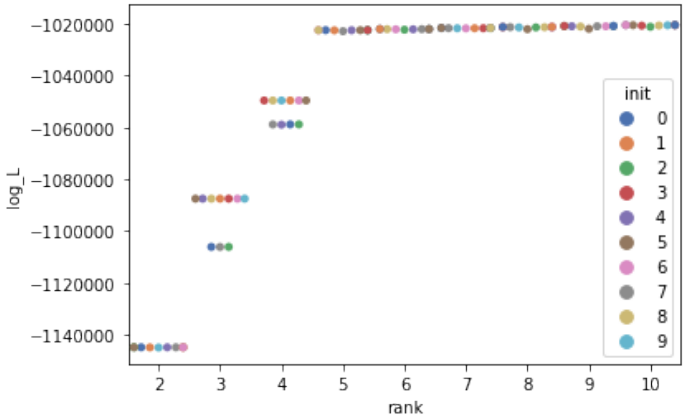
```
> experiment.data
{'/MyFirstExperiment/10', '/MyFirstExperiment/2', ..., \
'/MyFirstExperiment/9'}
```

The `Experiment` class computes a table of useful statistics,

```
> experiment.summary_table.head()
```

which, for example, enable us to inspect log likelihood of each initialisation[2],

```
> sns.swarmplot(x='rank', y='log_L', hue='init', \
data=experiment.summary_table, color='C0', palette='deep')
```

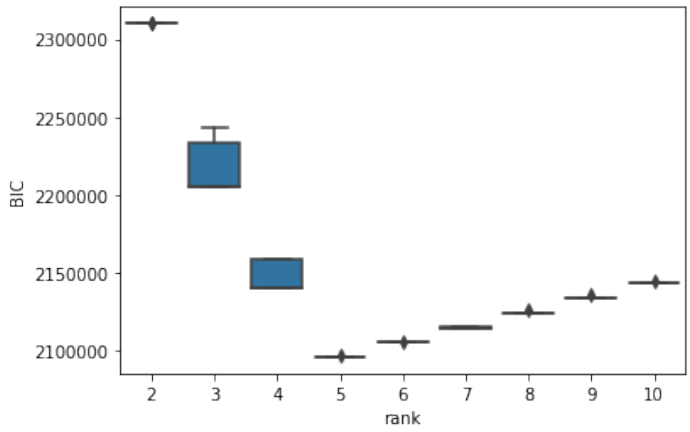


The `summary_table` also allows us to perform model selection using the Bayesian Information Criterion (BIC). This estimator tries to find a trade-off between the log-likelihood and the number of parameters in the model; chosen is the rank which minimises the BIC. To understand which model to choose in our experiment, we will quickly plot the rank against BIC,

```
> import seaborn as sns
> sns.boxplot(x='rank', y='BIC', data=experiment.summary_table, color='C0')
```

indicating that rank 5 is most appropriate for our dataset.

[2]Here we use the the `seaborn` library to create the plot. You can install the package, if necessary, by executing `pip install seaborn` in your terminal.



The `Cluster` class wraps multiple TensorSignature initialisations

We can extract the cluster of a specific decomposition rank by passing these keys to the getter function of an `Experiment` object. For example, to extract the rank 5 solution, we execute

```
> cluster = experiment['/MyFirstExperiment/5']
```

A `Cluster` instance is essentially a wrapper for multiple `Initializations` (Sec. A.3.4). It embodies attributes to access the parameters of a tensor signature inference, for example, we may access the extracted signature tensor(s) through the `S` field of `Cluster`.

```
> cluster.S.shape
(3, 3, 3, 5, 96, 5, 10)
```

Note the similarity between the shape of the extracted signature tensor and the shape of the input `snv` count tensor (`(3, 3, 3, 5, 96, 100)`). First few indices match the size of corresponding genomic dimensions, i.e. transcription and replication directionality (each `3`), genomic dimension 1 and 2 (`3` and `5`) and single base substitution types (`96`). The following two indices, however, indicate the decomposition rank (`5`) rather than the number of samples, and the number initialisations in the cluster.

Other model parameters may be accessed through the following fields:

- Other mutation type signatures: `result.T`
- Exposures: `result.E`
- Transcription and replicational strand biases: `result.b`
- Signature activities in transcribed/untranscribed regions and early/late replicating regions: `result.a`

- Arbitrary genomic property (like epigenetic signature activities): `result.k0`, `result.k1`, ..., `result.kx`
- Mixing proportions: `result.m`

The last dimension of an extracted Cluster parameter always indicates the number of available initialisations. To extract the solution of a particular initialisation, we can can simply index it using standard `numpy` indexing. Here we make use of the so called ellipsis operator (`...`) which enables to index the last dimension of an multidimensional array

```
> solution = cluster.b[..., 3]
> solution.shape
(2, 7)
```

`Cluster` objects provide an `init` field containing the index of the initialisation with the highest log-likelihood. To extract this particular `Initialization` from a cluster, we simply pass it to the `Cluster` getter function.

```
> init = cluster[cluster.init]
```

A `Initialization` object stores tensor signatures, factors and exposures

Similar to TensorSignature `Cluster` objects, `Initializations` contain the fitted model including all parameters. For example, we can access the extracted signature tensor by accessing the `S` field from `init`.

```
> init.S.shape
(3, 3, 3, 5, 96, 5, 1)
```

Note that the last dimension of `S` has a size of one, indicating an initialisation rather than a clustered signature tensor. Two other useful methods of `Initialization` objects are `to_dic` and `dump`, which let us serialise and save the result of a TensorSignature initialisation to the hard disk.

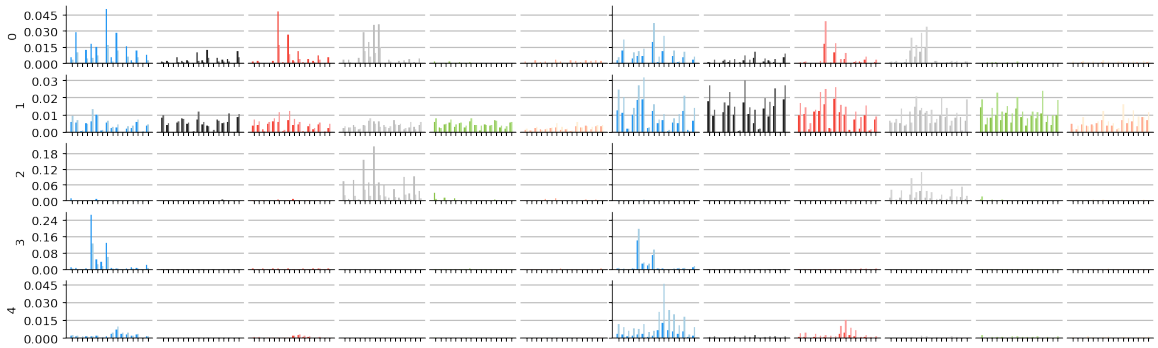
```
> # returns a dictionary with all parameters
> init.to_dic()
> # saves the initialisation to disk (load a saved solution with \
ts.load_dump)
> init.dump('initialisation.pkl')
```

The TensorSignatures API features some basic plotting function which allow us to visualise the extracted parameters of an `Initialisation`.

- `plot_signatures`: plots single base substitution spectra in context of transcription and replication
- `heatmap`: plots tensor factors (transcription and replication biases (`b`), signature activities (`a`), and genomic activities (`k0`, `k1`, ..., `kx`))

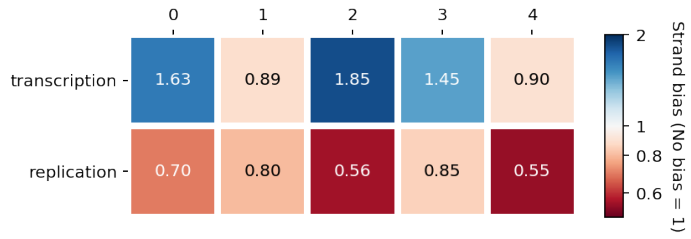
The `ts.plot_signatures` function expects an 5 dimensional array `(3, 3, -1, 96, rank)`. Due to the fact that we can have an arbitrary number of genomic states, we first have to reshape the signature tensor before we can pass it to the plotting function.

```
> plt.figure(figsize=(16, 5))
> ts.plot_signatures(init.S.reshape(3, 3, -1, 96, init.rank))
```



We can plot extracted tensor factors `result.b`, `result.a`, `result.k0` and `result.k1` using the `ts.heatmap` function. Note, that similarly to the the signature tensor, the `Initialization` object appends an additional dimension to indicate its index. For this reason, we need to reshape the arrays containing tensor factors or index them appropriately.

```
> result.b.shape # transcription and replication strand biases
(2, 5, 1)
> # ... (the elipsis operator) allows to index the last dimension of an array
> plt.figure(figsize=(6, 2))
> ts.heatmap(result.b[..., 0],
             vmin=.5, vmax=2,
             row_labels=['transcription', 'replication'],
             col_labels=['{}'.format(i) for i in range(5)],
             cbarlabel='Strand bias (No bias = 1)' # color bar label
            )
```



Exercise: Compare how extracted parameters differ from the ground truth (Sec. A.3.2).

Running TensorSignatures through the API

In some scenarios it might be desirable to run TensorSignatures via the API rather than the CLI (for example when integrating TensorSignatures into custom pipelines). To illustrate this, we first simulate data and extract the SNV count tensor and the matrix containing other mutation types. Here it is important to notice that the sample dimensions have to match, e.g. `snv[..., 4]` has to match `other[..., 4]`.

```
> data = ts.TensorSignatureData(
    seed=573, # set a seed for reproducibility
    rank=5, # number of signatures
    samples=100, # number of samples
    dimensions=[3, 5], # number of arbitrary genomic dimensions
    mutations=1000)
> snv = data_set.snv() # the SNV count tensor (3, 3, 3, 5, 96, 100)
> other = data_set.other() # other mutation type matrix (234, 100)
```

The next step is to pass the desired decomposition rank, as well as the input data, i.e. the `snv` count tensor and the `other` mutation matrix, to the `TensorSignature` class[3]. The `TensorSignature` constructor also receives other model hyperparameters such as learning rate of the model or the number of epochs to train the model. By default, TensorSignatures uses the outlier robust negative binomial distribution with a dispersion $\tau = 50$ to model the mutation count, and trains the model for 10,000 epochs.

```
# perform a rank 5 decomposition
model = ts.TensorSignature(snv, other, rank=5, verbose=True, epochs=20000)
```

[3]When working with real genomic data it is also necessary to pass a normalisation tensor (via the argument `N`) which accounts for differences in the nucleotide composition of different genomic regions to the `TensorSignature` constructor.

To fit the signatures to our data, we simply invoke the `fit` method of the `model` instance, which will return a `Initialization` object after finishing to train the model.

```
> result = model.fit()
```

Appendix B

Supplementary Figures

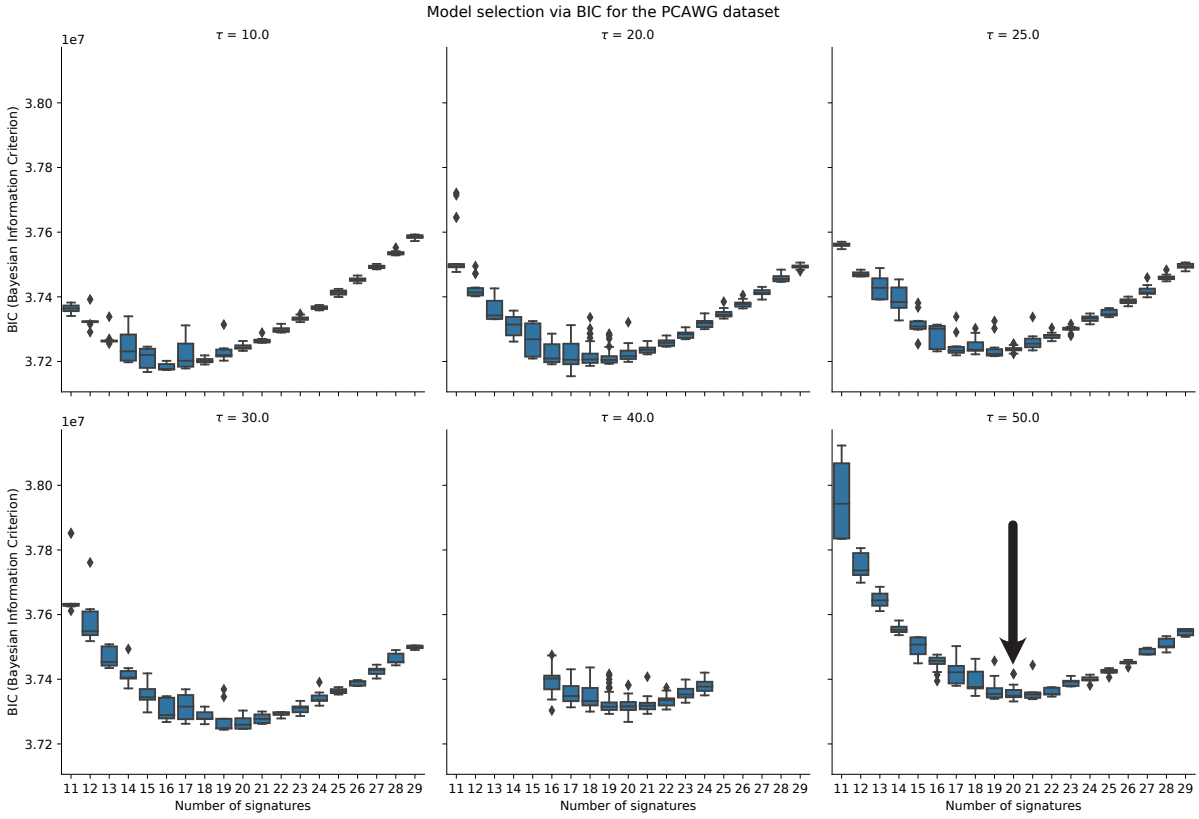


Fig. B.1 **Model selection in the PCAWG dataset.** Chosen number of signatures 20 with a size τ of 50.

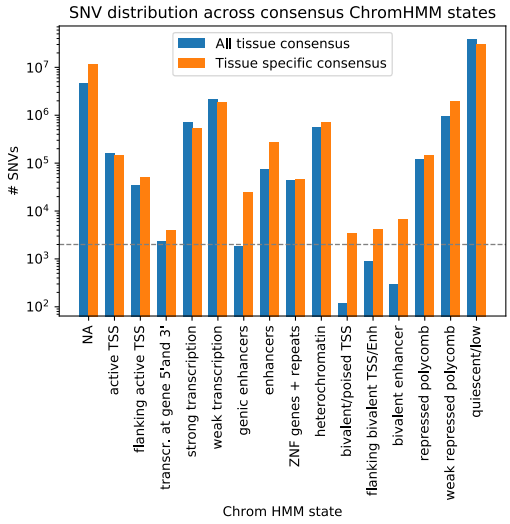


Fig. B.2 **Distribution of PCAWG SNV count data across Chrom-HMM states using an all tissue and tissue specific consensus.**

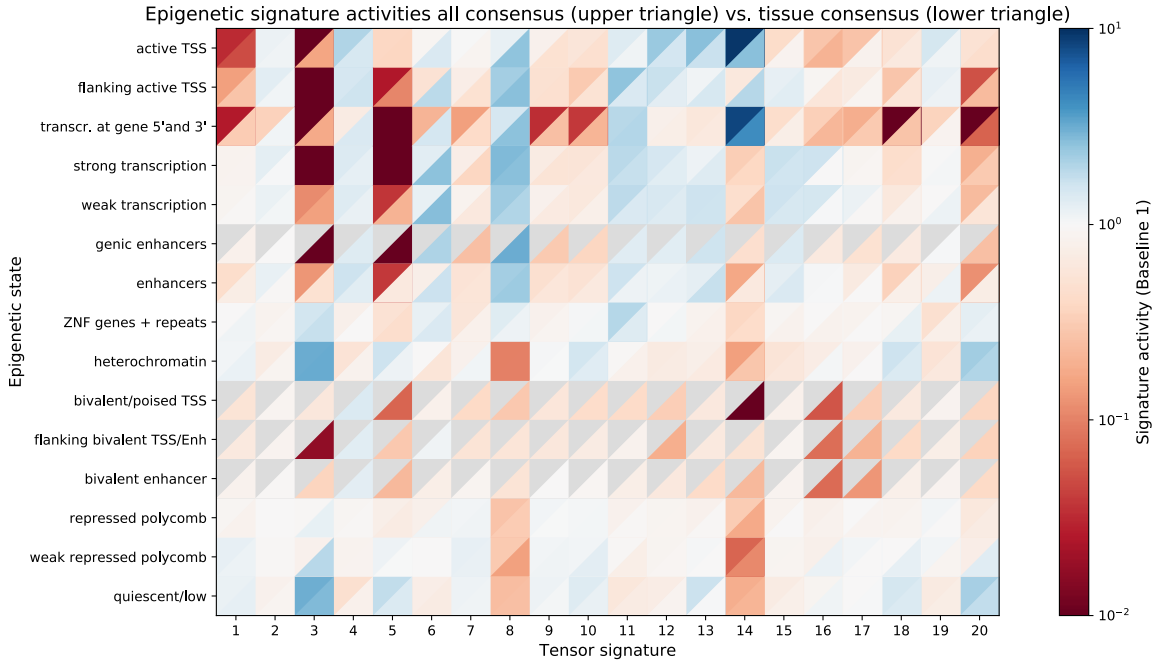


Fig. B.3 **Annotating SNVs with consensus and partially matched ChromHMM states.** Comparison of inferred epigenetic signature activities using an all tissue (upper triangle) and tissue specific (lower triangle) consensus (grey triangles indicate NA values).

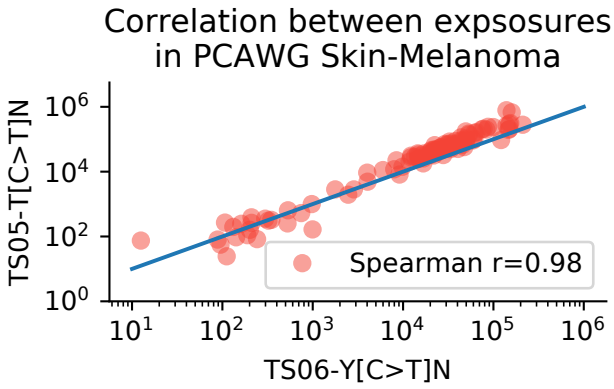


Fig. B.4 **Correlation of TS05 and TS06 exposures in Skin-Melanoma samples.** Blue line indicates the identity line ($y = x$).

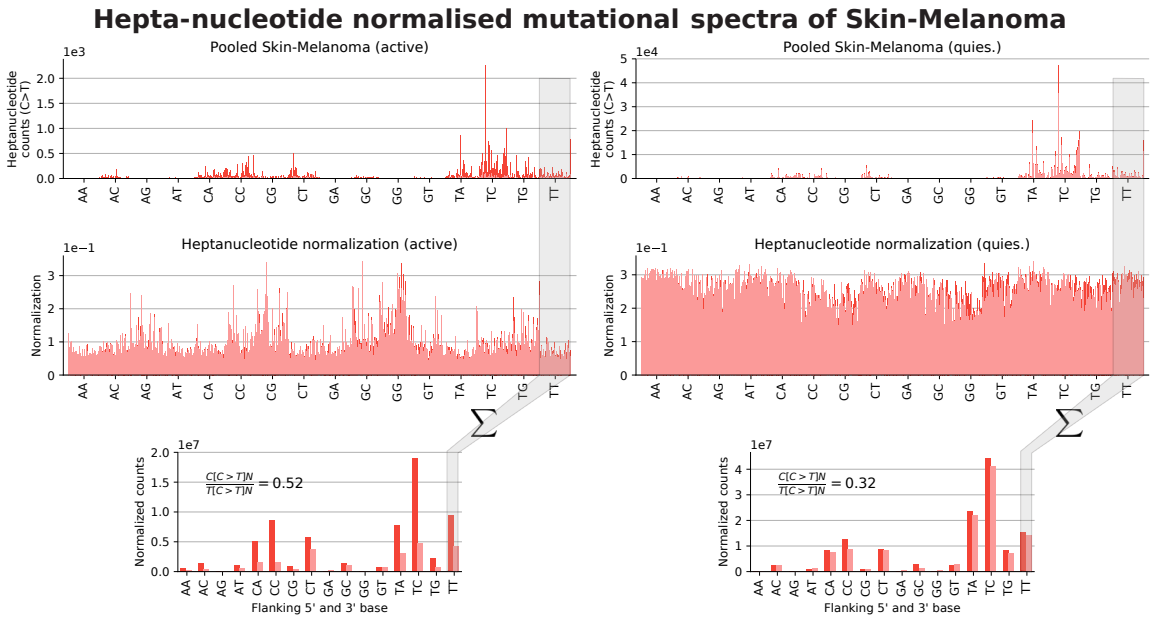


Fig. B.5 **Heptanucleotide context normalised C>T mutation counts in active and quiescent genomic regions.**

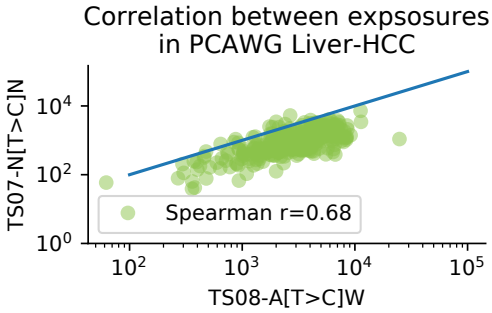


Fig. B.6 **Correlation of predicted TS07 and TS08 mutation counts in Liver-HCC samples.** Blue line indicates the identity line ($y = x$).

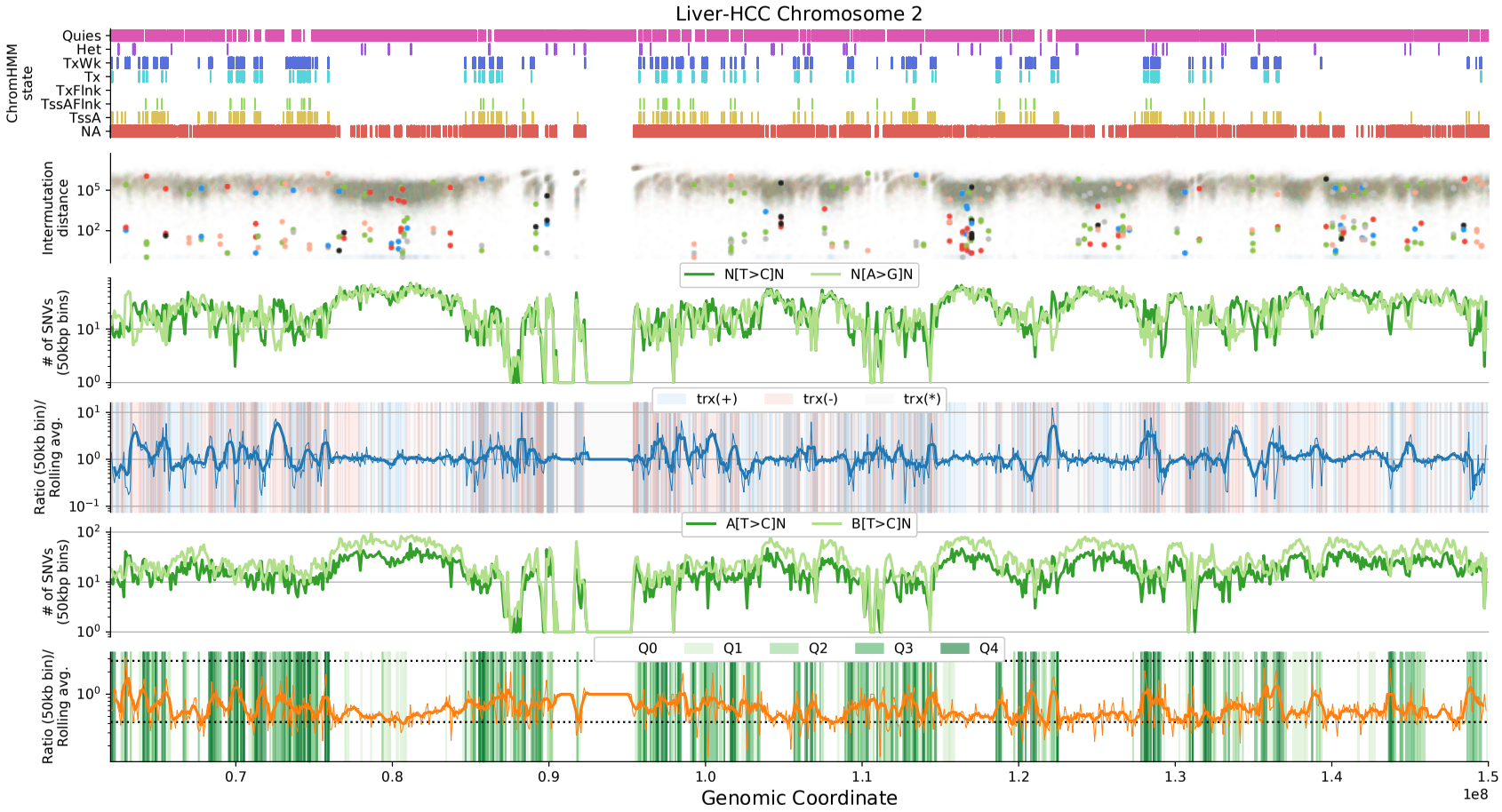


Fig. B.7 **The spatial distribution of T>C mutations in Liver-HCC.**. Upper panel: Consensus ChromHMM states from chromosome 2 depicting several active and quiescent genomic region, and the corresponding mutational density from pooled Liver-HCC samples. Middle panel: Illustration of the transcriptional strand bias in terms of 100kbp binned N[T>C]N and N[A>G]N counts, and respective ratio (thin blue line). The thick blue line depicts the corresponding rolling average over 5 consecutive bins. Lower panel: Changes in the distribution of T>C mutations in an active and quiescent genomic regions in terms of 100kpb binned A[T>C]N and B[T>C]N counts. Thin orange line: A[T>C]/B[T>C] ratio, thick orange line: rolling average over 5 consecutive bins.

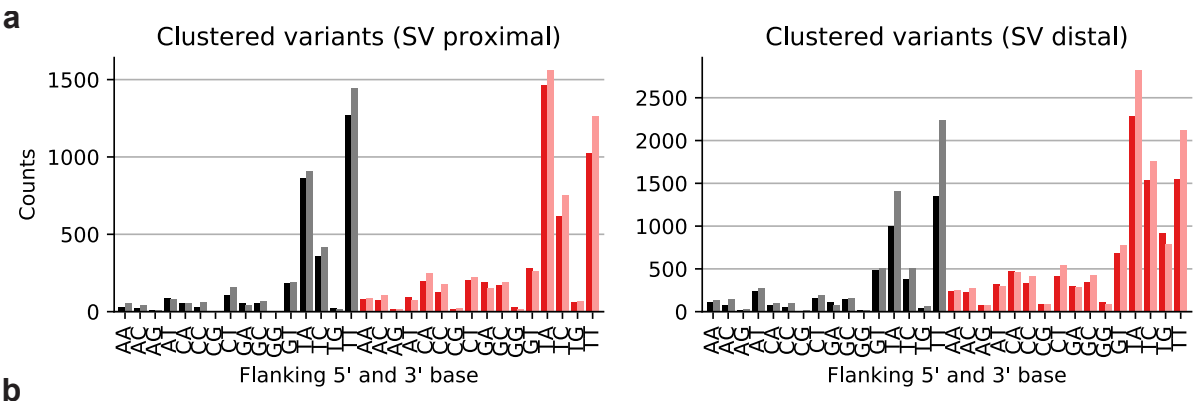


Fig. B.8 **Pancancer-wide pooled C>G and C>T clustered variants proximal and distal to SVs.**

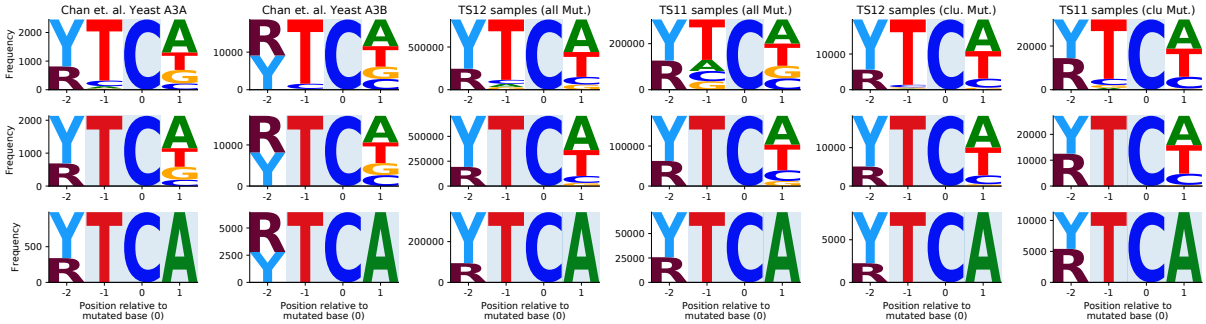


Fig. B.9 **Tetranucleotide motifs at sites of APOBEC mutations.**

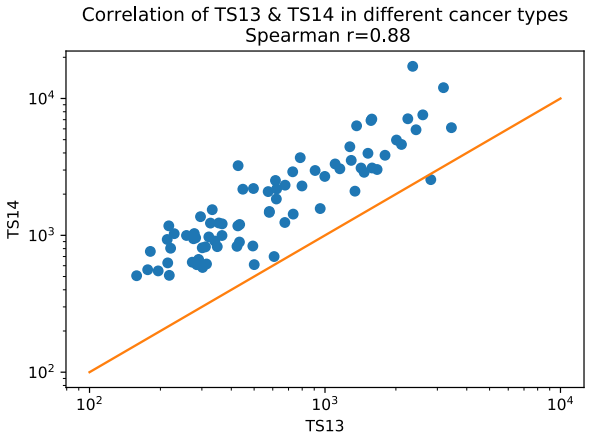


Fig. B.10 **Correlation of TS13 and TS14 exposures in lymphoid cancers (Lymph-BNHL/CLL/NOS).**

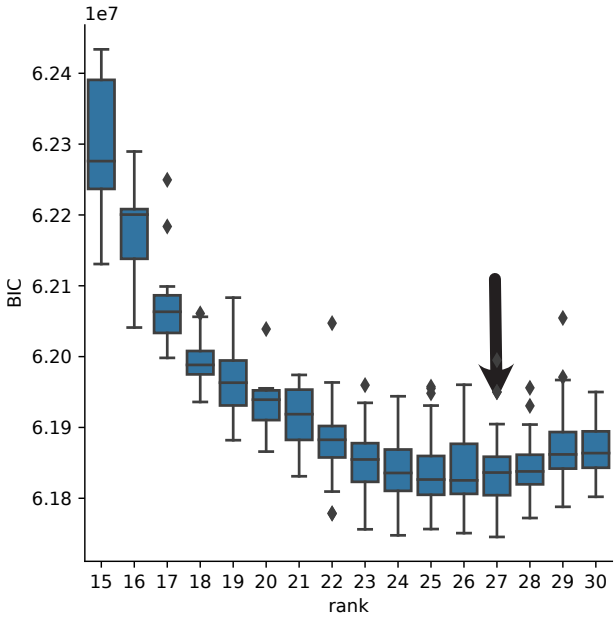


Fig. B.11 **Model selection in the HMF dataset (chosen number of signatures 27 with a size of 30).**

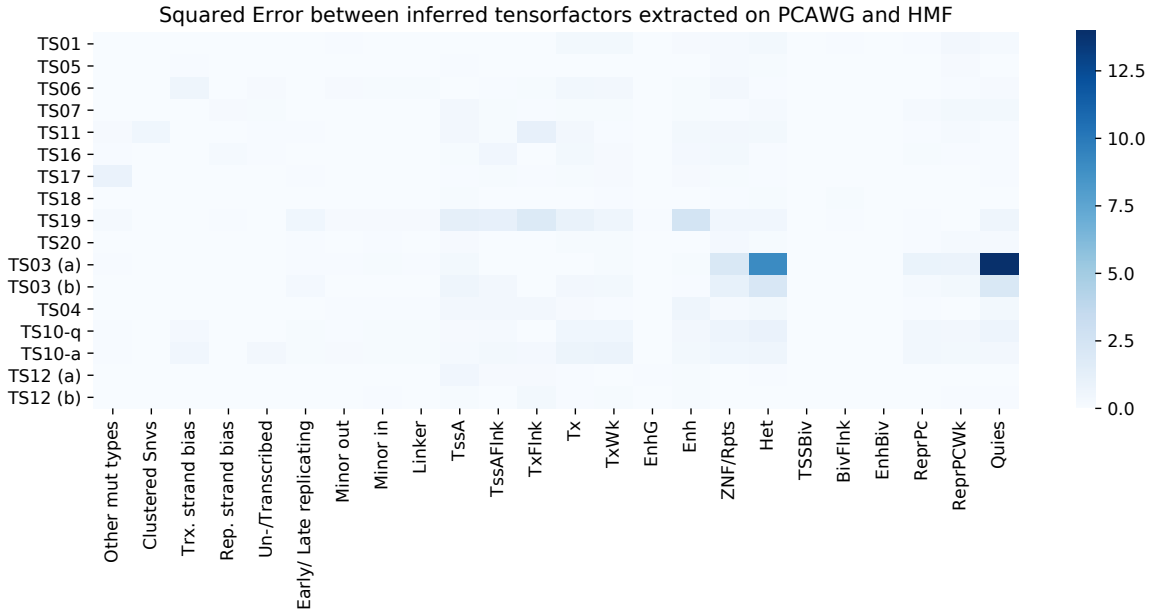


Fig. B.12 **Squared errors of tensor factors from the PCAWG discovery and HMF validation analysis.**

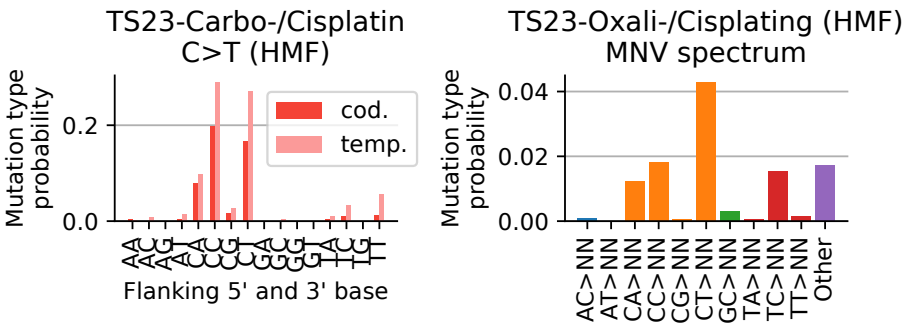


Fig. B.13 **C>T mutation type probabilities of TS22 for coding and template strand DNA, and the MNV spectrum of TS23.**

Table B.1 Features of the SNV count tensor

index	feature	index	feature	index	feature
0	A[C>A]A	32	A[C>T]A	64	A[T>C]A
1	A[C>A]C	33	A[C>T]C	65	A[T>C]C
2	A[C>A]G	34	A[C>T]G	66	A[T>C]G
3	A[C>A]T	35	A[C>T]T	67	A[T>C]T
4	C[C>A]A	36	C[C>T]A	68	C[T>C]A
5	C[C>A]C	37	C[C>T]C	69	C[T>C]C
6	C[C>A]G	38	C[C>T]G	70	C[T>C]G
7	C[C>A]T	39	C[C>T]T	71	C[T>C]T
8	G[C>A]A	40	G[C>T]A	72	G[T>C]A
9	G[C>A]C	41	G[C>T]C	73	G[T>C]C
10	G[C>A]G	42	G[C>T]G	74	G[T>C]G
11	G[C>A]T	43	G[C>T]T	75	G[T>C]T
12	T[C>A]A	44	T[C>T]A	76	T[T>C]A
13	T[C>A]C	45	T[C>T]C	77	T[T>C]C
14	T[C>A]G	46	T[C>T]G	78	T[T>C]G
15	T[C>A]T	47	T[C>T]T	79	T[T>C]T
16	A[C>G]A	48	A[T>A]A	80	A[T>G]A
17	A[C>G]C	49	A[T>A]C	81	A[T>G]C
18	A[C>G]G	50	A[T>A]G	82	A[T>G]G
19	A[C>G]T	51	A[T>A]T	83	A[T>G]T
20	C[C>G]A	52	C[T>A]A	84	C[T>G]A
21	C[C>G]C	53	C[T>A]C	85	C[T>G]C
22	C[C>G]G	54	C[T>A]G	86	C[T>G]G
23	C[C>G]T	55	C[T>A]T	87	C[T>G]T
24	G[C>G]A	56	G[T>A]A	88	G[T>G]A
25	G[C>G]C	57	G[T>A]C	89	G[T>G]C
26	G[C>G]G	58	G[T>A]G	90	G[T>G]G
27	G[C>G]T	59	G[T>A]T	91	G[T>G]T
28	T[C>G]A	60	T[T>A]A	92	T[T>G]A
29	T[C>G]C	61	T[T>A]C	93	T[T>G]C
30	T[C>G]G	62	T[T>A]G	94	T[T>G]G
31	T[C>G]T	63	T[T>A]T	95	T[T>G]T

Table B.2 MNV features of the other mutation count matrix.

i	type	feature	i	type	feature	i	type	feature
0	MNV	CA>AC	30	MNV	CT>AC	60	MNV	TT>AG
1	MNV	CA>GC	31	MNV	CT>GC	61	MNV	TT>CG
2	MNV	CA>TC	32	MNV	CT>TC	62	MNV	TT>GG
3	MNV	CA>AG	33	MNV	CT>AG	63	MNV	AC>CA
4	MNV	CA>GG	34	MNV	CT>GG	64	MNV	AC>GA
5	MNV	CA>TG	35	MNV	CT>TG	65	MNV	AC>TA
6	MNV	CA>AT	36	MNV	TA>AC	66	MNV	AC>CG
7	MNV	CA>GT	37	MNV	TA>CC	67	MNV	AC>GG
8	MNV	CA>TT	38	MNV	TA>GC	68	MNV	AC>TG
9	MNV	CC>AA	39	MNV	TA>AG	69	MNV	AC>CT
10	MNV	CC>GA	40	MNV	TA>CG	70	MNV	AC>GT
11	MNV	CC>TA	41	MNV	TA>GG	71	MNV	AC>TT
12	MNV	CC>AG	42	MNV	TA>AT	72	MNV	AT>CA
13	MNV	CC>GG	43	MNV	TA>CT	73	MNV	AT>GA
14	MNV	CC>TG	44	MNV	TA>GT	74	MNV	AT>TA
15	MNV	CC>AT	45	MNV	TC>AA	75	MNV	AT>CC
16	MNV	CC>GT	46	MNV	TC>CA	76	MNV	AT>GC
17	MNV	CC>TT	47	MNV	TC>GA	77	MNV	AT>TC
18	MNV	CG>AA	48	MNV	TC>AG	78	MNV	AT>CG
19	MNV	CG>GA	49	MNV	TC>CG	79	MNV	AT>GG
20	MNV	CG>TA	50	MNV	TC>GG	80	MNV	AT>TG
21	MNV	CG>AC	51	MNV	TC>AT	81	MNV	GC>AA
22	MNV	CG>GC	52	MNV	TC>CT	82	MNV	GC>CA
23	MNV	CG>TC	53	MNV	TC>GT	83	MNV	GC>TA
24	MNV	CG>AT	54	MNV	TT>AA	84	MNV	GC>AG
25	MNV	CG>GT	55	MNV	TT>CA	85	MNV	GC>CG
26	MNV	CG>TT	56	MNV	TT>GA	86	MNV	GC>TG
27	MNV	CT>AA	57	MNV	TT>AC	87	MNV	GC>AT
28	MNV	CT>GA	58	MNV	TT>CC	88	MNV	GC>CT
29	MNV	CT>TA	59	MNV	TT>GC	89	MNV	GC>TT
⋮	⋮	⋮	⋮	⋮	⋮	90	MNV	MNV(other)

Table B.3 Indel features of the other mutation count matrix.

index	type	feature	index	type	feature
91	Indel	delC	122	Indel	insT
92	Indel	delT	123	Indel	insCA
93	Indel	delCA	124	Indel	insCC
94	Indel	delCC	125	Indel	insCG
95	Indel	delCG	126	Indel	insCT
96	Indel	delCT	127	Indel	insTA
97	Indel	delTA	128	Indel	insTC
98	Indel	delTC	129	Indel	insTT
99	Indel	delTT	130	Indel	insAC
100	Indel	delAC	131	Indel	insAT
101	Indel	delAT	132	Indel	insGC
102	Indel	delGC	133	Indel	ins3
103	Indel	del3	134	Indel	ins4
104	Indel	del4	135	Indel	ins5
105	Indel	del5	136	Indel	ins6
106	Indel	del6	137	Indel	ins7
107	Indel	del7	138	Indel	ins8
108	Indel	del8	139	Indel	ins9
109	Indel	del9	140	Indel	ins10
110	Indel	del10	141	Indel	ins(10,20]
111	Indel	del(10,20]	142	Indel	ins(20,30]
112	Indel	del(20,30]	143	Indel	ins(30,40]
113	Indel	del(30,40]	144	Indel	ins(40,50]
114	Indel	del(40,50]	145	Indel	ins(50,60]
115	Indel	del(50,60]	146	Indel	ins(60,70]
116	Indel	del(60,70]	147	Indel	ins(70,80]
117	Indel	del(70,80]	148	Indel	ins(80,90]
118	Indel	del(80,90]	149	Indel	ins(90,100]
119	Indel	del(90,100]	150	Indel	ins(100,Inf]
120	Indel	del(100,Inf]	151	Indel	indel
121	Indel	insC	152	Indel	indel(other)

Table B.4 SV features of the other mutation count matrix.

index	type	feature
153	SV	del:0-1e4
154	SV	del:1e4-3e6
155	SV	del:3e6+
156	SV	fb
157	SV	fragile_site_del
158	SV	fragile_site_td
159	SV	inv
160	SV	other_unbal
161	SV	simple_unbal
162	SV	td:0-5.5e4
163	SV	td:1e7+
164	SV	td:2e6-1e7
165	SV	td:5.5e4-2e6
166	SV	A+/B+
167	SV	A+^B+
168	SV	bal_bkpt:0-1e2
169	SV	bal_bkpt:1e2-1e5
170	SV	bal_bkpt:1e5+
171	SV	bal_transloc:0-1e2
172	SV	bal_transloc:1e2-1e5
173	SV	bal_transloc:1e5+
174	SV	chromoplexy_chain:0-1e2
175	SV	chromoplexy_chain:1e2-1e5
176	SV	chromoplexy_chain:1e5+
177	SV	chromoplexy_cycle:0-1e2
178	SV	chromoplexy_cycle:1e2-1e5
179	SV	chromoplexy_cycle:1e5+
180	SV	shard_chain:0-1e3
181	SV	shard_chain:1e3-1e5
182	SV	shard_chain:1e5+
183	SV	shard_cycle:0-1e3
184	SV	shard_cycle:1e3-1e5
185	SV	shard_cycle:1e5+
186	SV	TD_after_unbal_transloc
187	SV	A+/B+/C+
188	SV	A+/B+/D-
189	SV	A+/B+^C+
190	SV	A+/C-/C+
191	SV	A+^B+/C+
192	SV	A+^B+/D-
⋮	⋮	⋮

index	type	feature
193	SV	A+^C+/C-
194	SV	A+^D-/B+
195	SV	B-/B+/C+
196	SV	B-^C-/C+
197	SV	B-^C+/B+
198	SV	A+/B+/D-/D+
199	SV	A+/C-/C+/D+
200	SV	A+/C-/C+/E-
201	SV	A+/C-/D-/D+
202	SV	A+^B+/C+^D+
203	SV	A+^B+/D-^E-
204	SV	A+^C+/B+^E-
205	SV	A+^C+/C-/E-
206	SV	A+^C+/C-^D+
207	SV	A+^C+/C-^E-
208	SV	A+^D-/B+^D+
209	SV	A+^D+/B+^D-
210	SV	A+^E-/C-/C+
211	SV	B-/B+/C+/D+
212	SV	B-/B+/D-/D+
213	SV	B-/B+^D-/D+
214	SV	B-/C-/C+/D+
215	SV	B-^C-/C+/D+
216	SV	B-^C+/B+^D+
217	SV	B-^C+/C-/D+
218	SV	B-^D-/B+/D+
219	SV	B-^D+/B+/D-
220	SV	B-^D+/B+^C+
221	SV	B-^D+/C-/C+
222	SV	direct_inversion:0-1e5
223	SV	direct_inversion:1e5+
224	SV	dup_trp_dup
225	SV	fb_then_fb
226	SV	inversion_gain_loss:0-1e5
227	SV	inversion_gain_loss:1e5+
228	SV	inverted_duplication
229	SV	A+/B+/D-/D+/F-
230	SV	A+/C-/C+/D+/F-
231	SV	A+/C-/C+/E-/E+
232	SV	B-/B+/C+/E-/E+
233	SV	B-/B+/D-/D+/E+

Table B.5 TensorSignatures and equivalent SigProfiler (SBS) signatures.

type	feature
TS1	SBS1
TS2	no equivalent SBS
TS3	SBS40
TS4	SBS5
TS5	SBS7a
TS6	SBS7b
TS7	SBS12
TS8	SBS16
TS9	SBS22
TS10	SBS4
TS11	SBS2/13
TS12	SBS2/13
TS13	SBS89
TS14	SBS9
TS15	SBS6/15/26 + ID1/2
TS16	SBS14
TS17	SBS10a/b
TS18	SBS36
TS19	SBS3
TS20	SBS17a/b

Table B.6 A comparison of different mutational signature extraction tools.

Feature	TensorSignatures	SigProfiler	SparseSignatures	deconstructSigs	EMu	Mutalisk
Concurrent signature extraction across all variant types	Yes	No (only via separate NMFs)	No	No	No	No (does not perform denovo extraction)
Characterises extracted signatures with respect to various genomic properties	Yes (w.r.t. transcription, replication, epigenome, nucleosomal and clustering)	No (post-hoc approaches have been described)	No	No	Yes	No (but correlates variants with respect to different genomic features)
Noise robust signature extraction	Yes (using over-dispersed NB dist.)	No explicit noise modelling (conventional NMF or Poisson model)	Yes (using a lasso penalty and considers natural background noise of ‘standard’ replication error and repair)	NA (does not perform de novo signature extractions)	No explicit noise modelling (uses an equidispersed Poisson model)	NA (does not perform de novo signature extractions)
Fits exposures to set of predefined exposures	Yes	No	No	Yes	No	Yes (to up to seven signatures)
Provides a Webportal	Yes	No	No	No	No	Yes
Inference Method	Maximum likelihood estimation	Conventional NMF (iterative update rules)	Elaborate custom fitting procedure	Custom fitting procedure which includes the selection of appropriate signatures	Expectation-Maximization	Custom fitting procedure which requires the user to preselect appropriate signatures
Model selection approach	BIC	Combination of residuals (MSE) and signature stability (silhouette score)	Cross-validation	NA (does not perform de novo signature extractions)	BIC	NA (does not perform de novo signature extractions)

Appendix C

Vignettes

C.1 TS01-N[C>T]G (5meC>T)

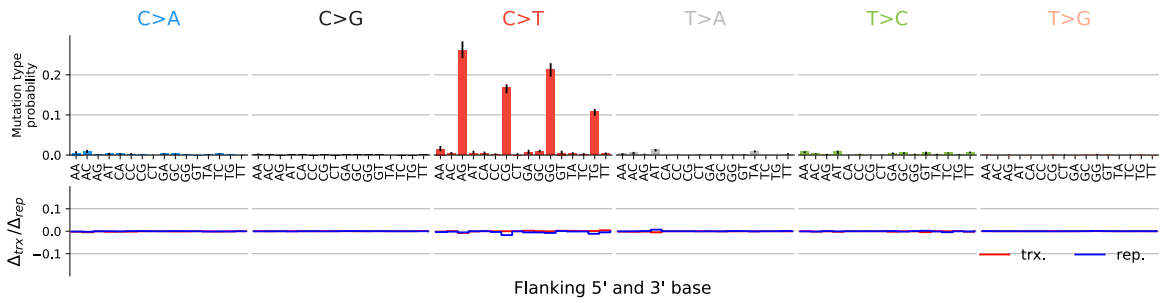


Fig. C.1 **TS01: Single base substitution spectrum.**

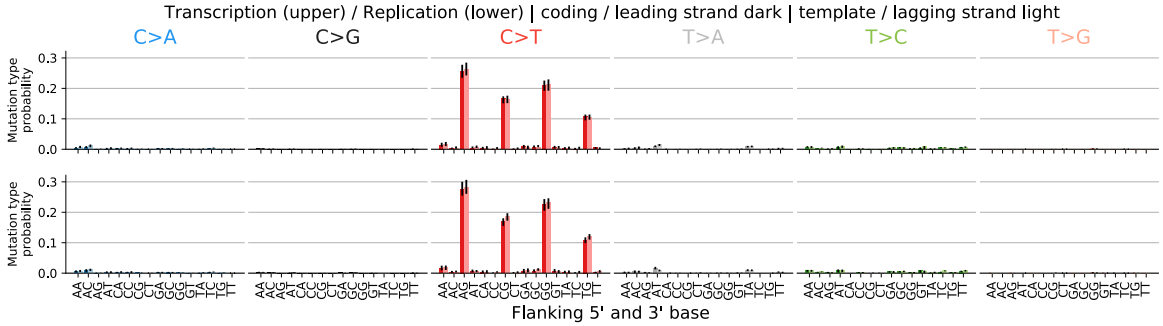


Fig. C.2 **TS01: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

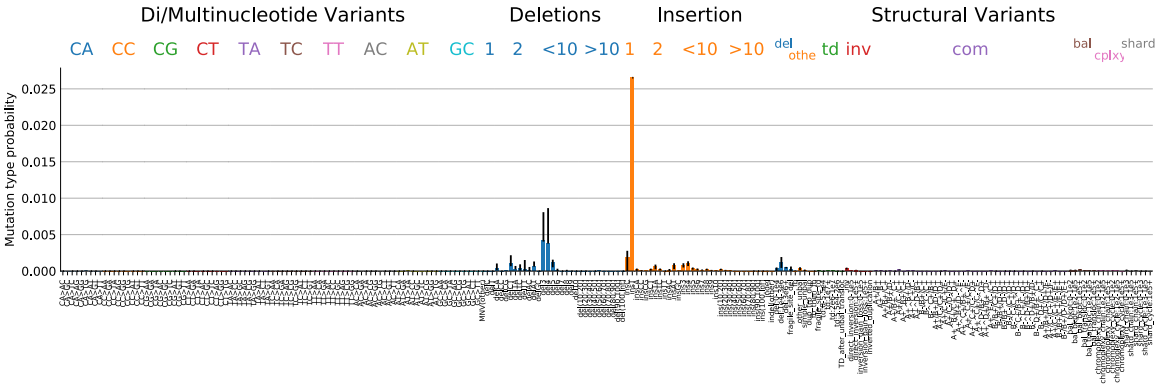


Fig. C.3 **TS01: Spectrum other mutation types.**

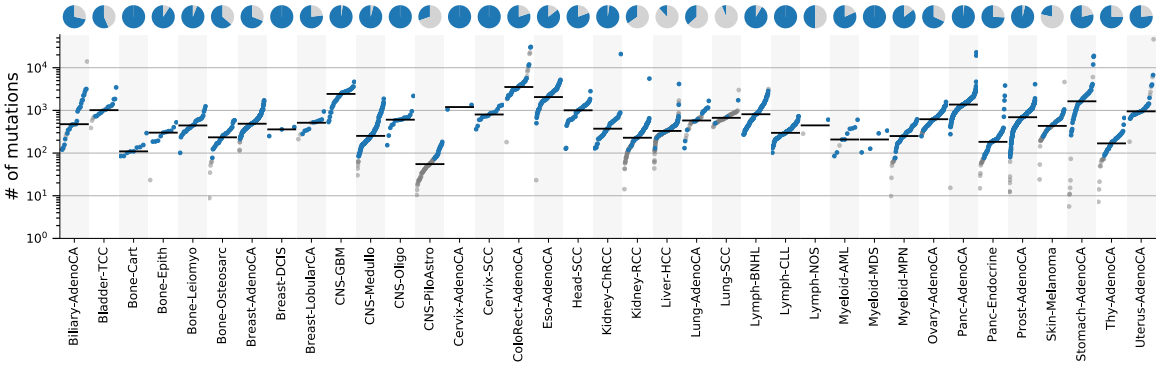


Fig. C.4 **TS01: Signature activity in different cancer types.**

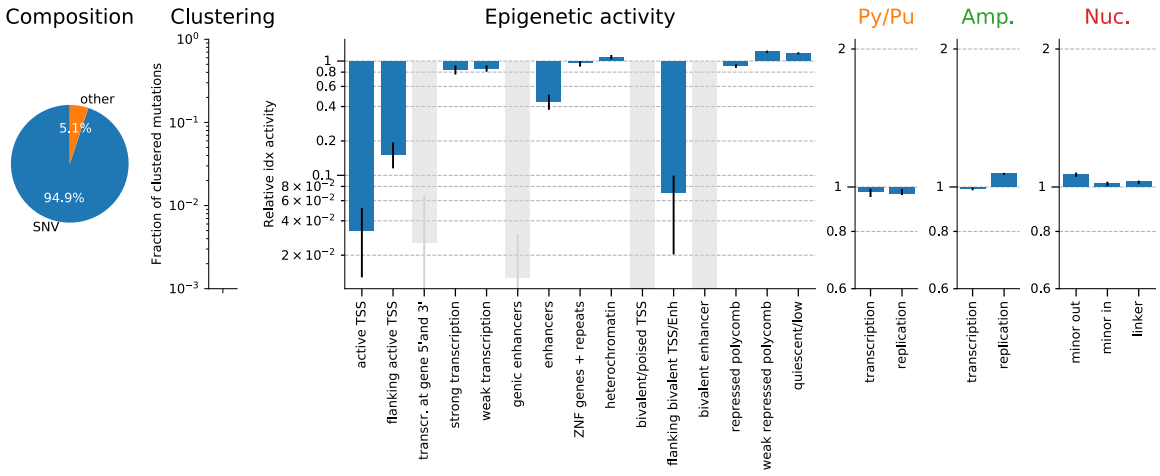


Fig. C.5 **TS01: Signature specific tensor coefficients.**

C.2 TS02-N[C>T]N (unknown)

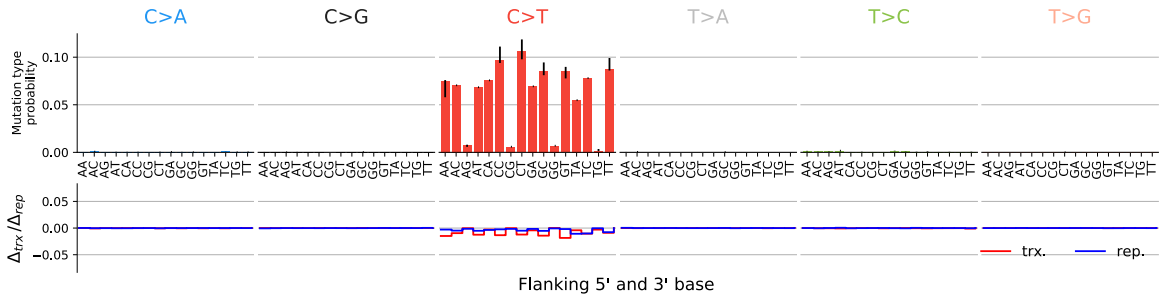


Fig. C.6 **TS02: Single base substitution spectrum.**

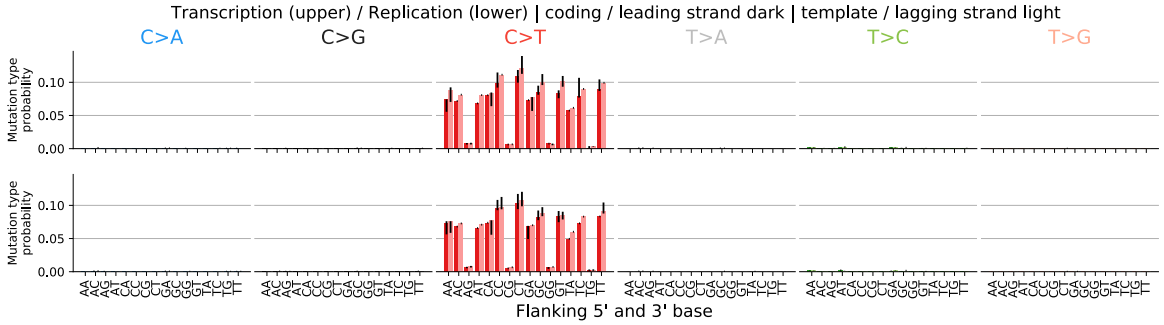


Fig. C.7 **TS02: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

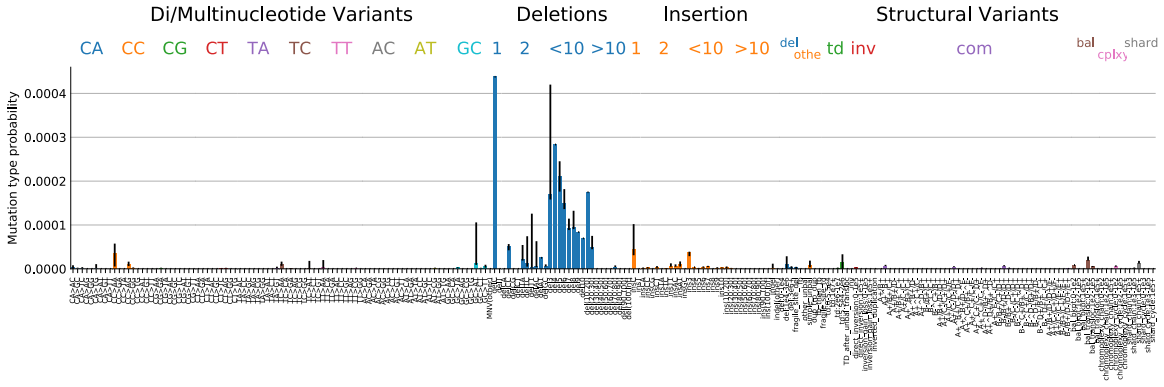


Fig. C.8 **TS02: Spectrum other mutation types.**

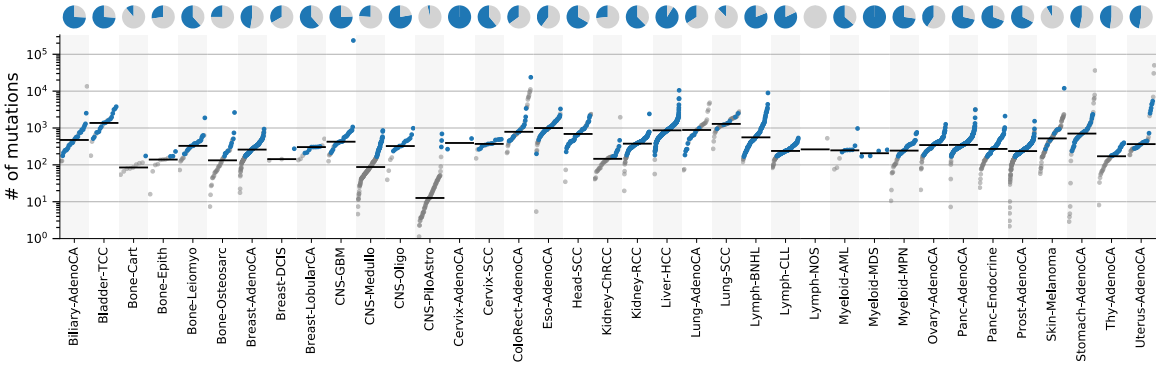


Fig. C.9 **TS02: Signature activity in different cancer types.**

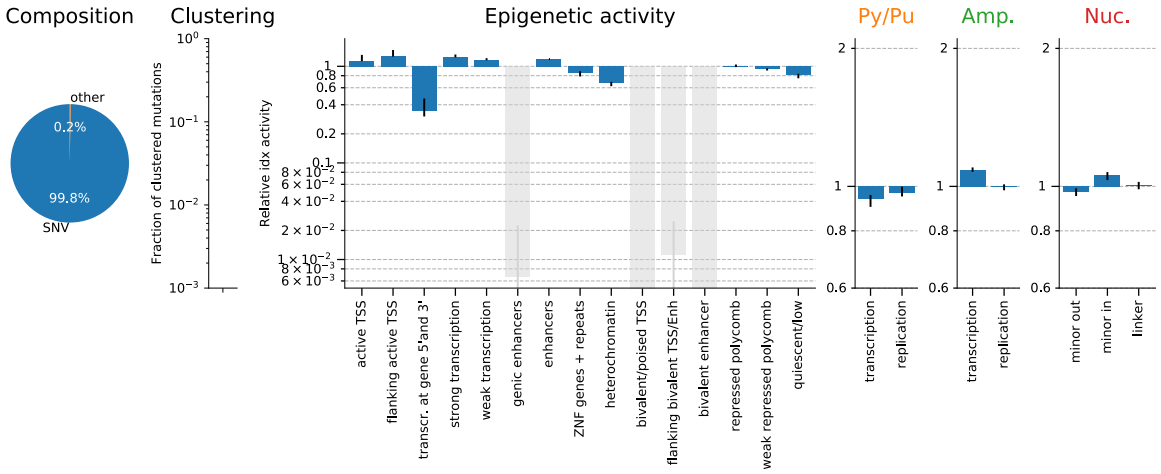


Fig. C.10 **TS02: Signature specific tensor coefficients.**

C.3 TS03-N[N>N]N-q (unknown/quiet)

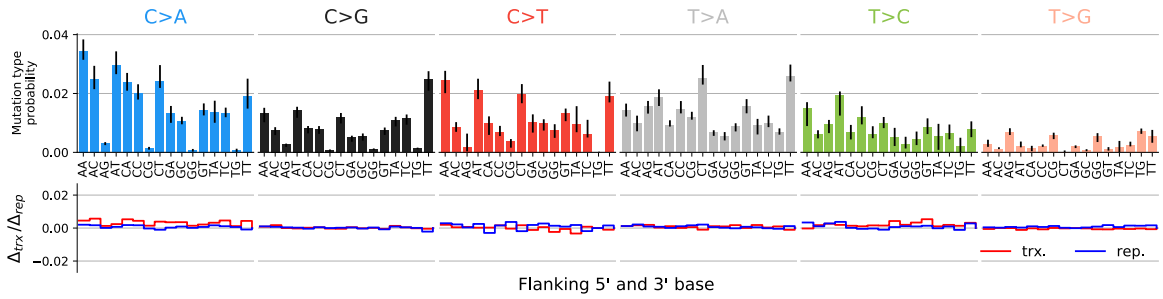


Fig. C.11 **TS03: Single base substitution spectrum.**

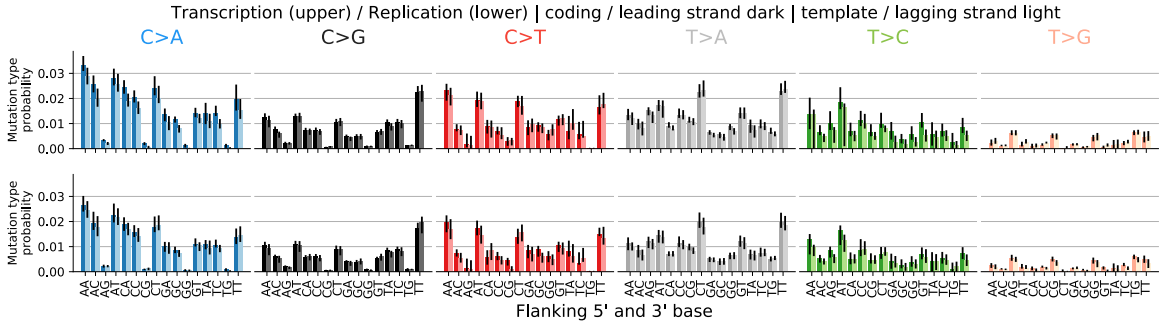


Fig. C.12 **TS03: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

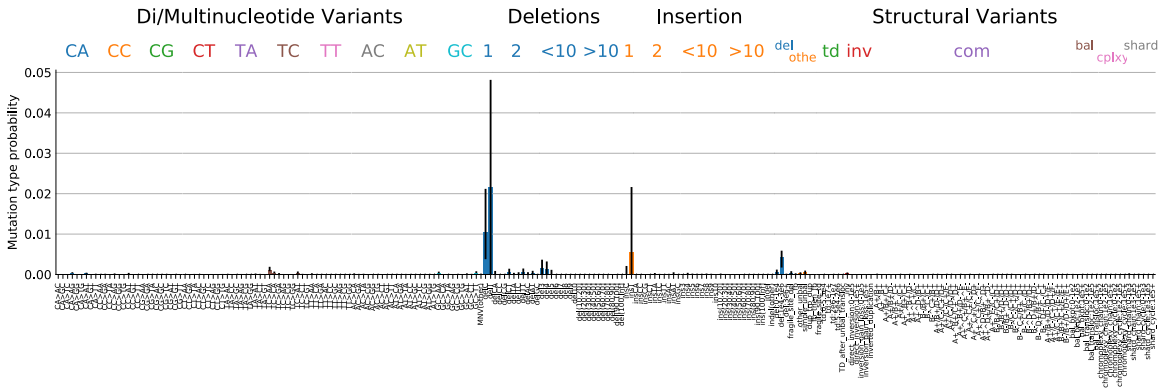


Fig. C.13 **TS03: Spectrum other mutation types.**

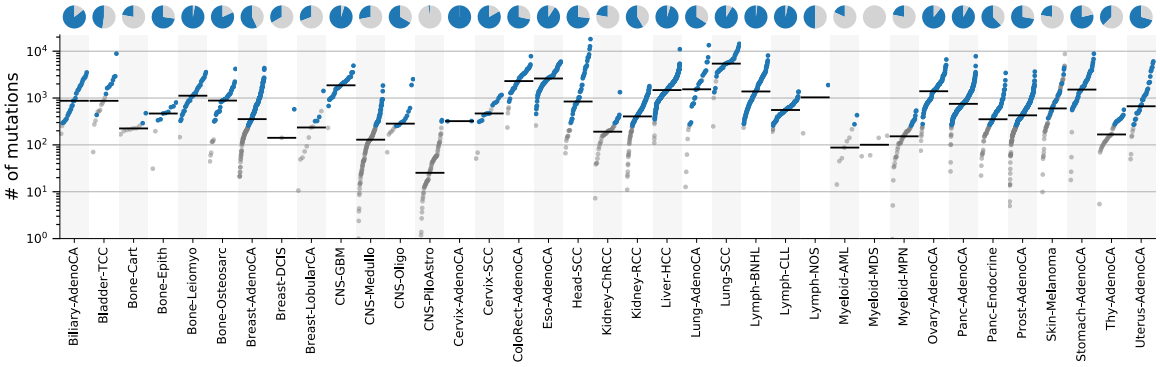


Fig. C.14 **TS03: Signature activity in different cancer types.**

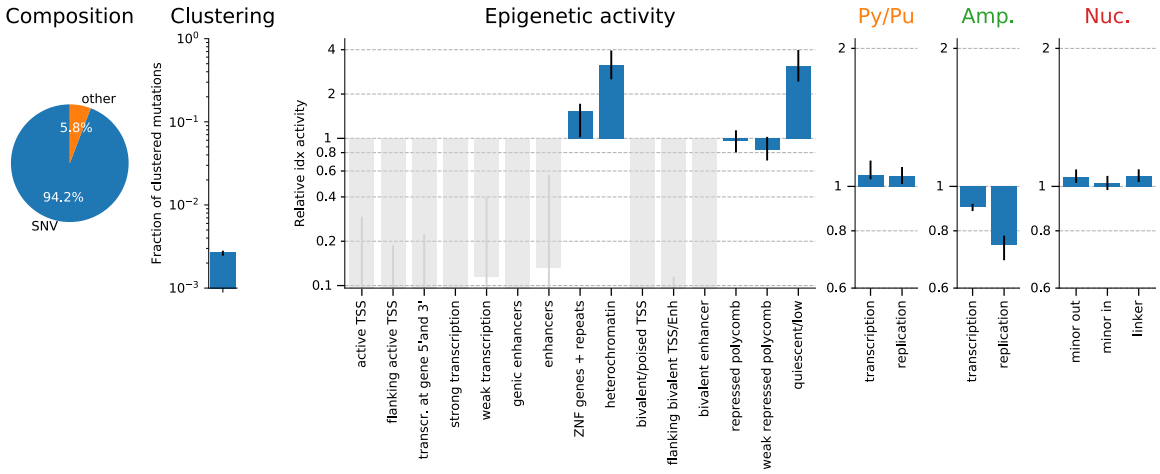


Fig. C.15 **TS03: Signature specific tensor coefficients.**

C.4 TS04-N[N>N]N (unknown/active)

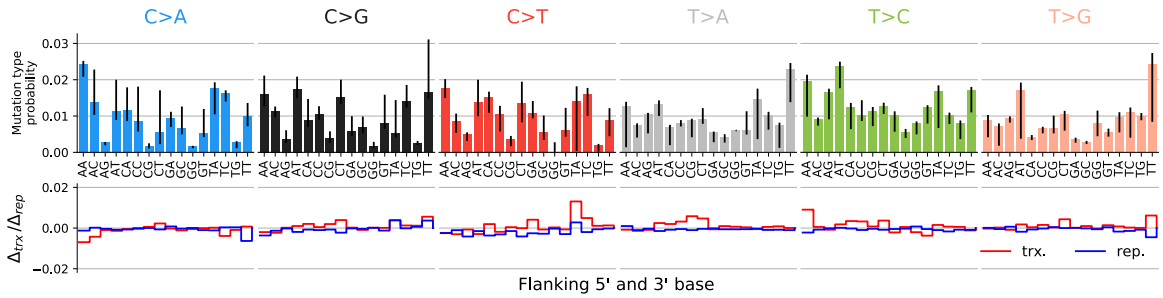


Fig. C.16 **TS04: Single base substitution spectrum.**

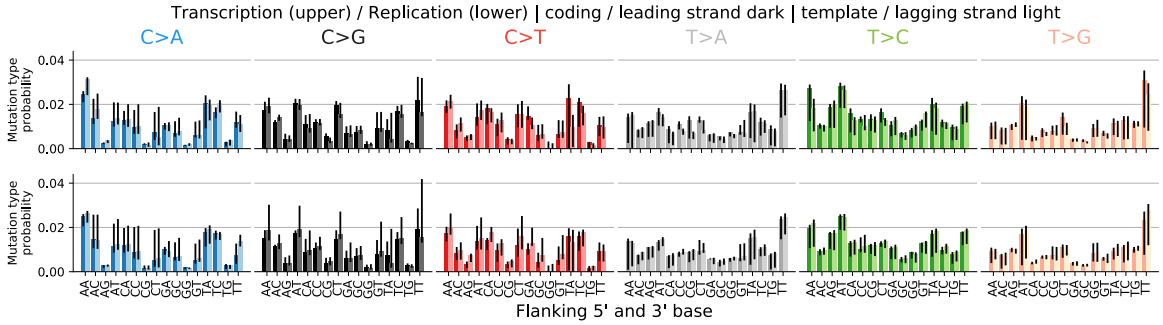


Fig. C.17 **TS04: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

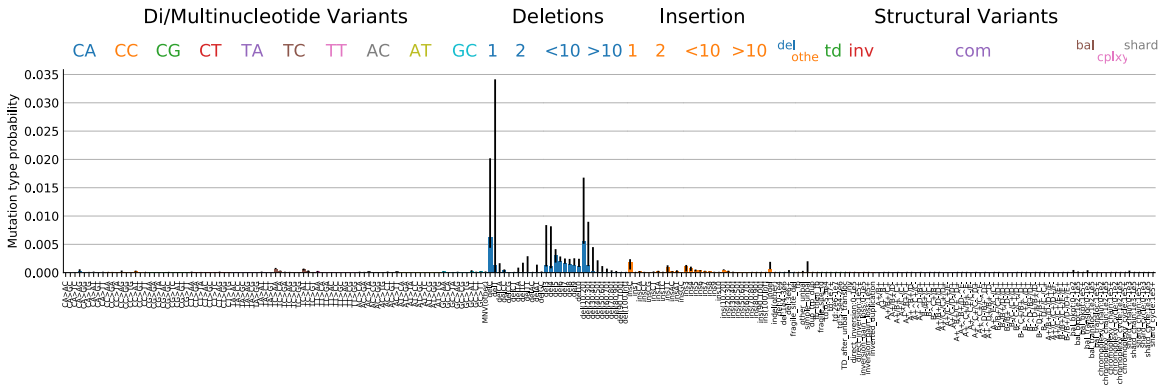


Fig. C.18 **TS04: Spectrum other mutation types.**

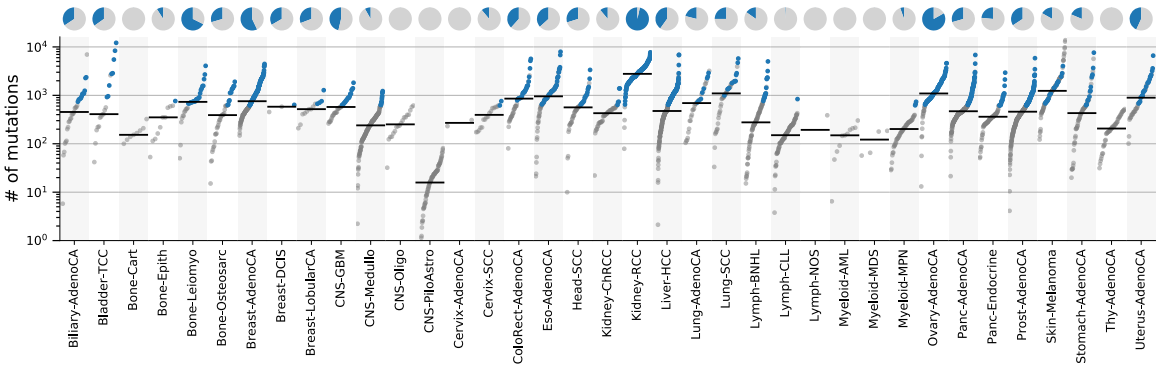


Fig. C.19 **TS04: Signature activity in different cancer types.**

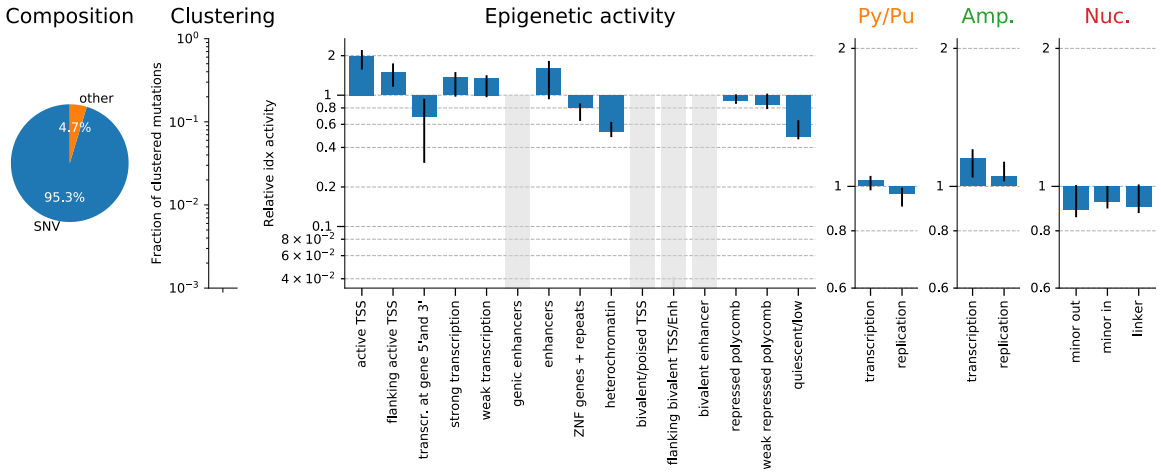


Fig. C.20 **TS04: Signature specific tensor coefficients.**

C.5 TS05-T[C>T]N (UV/GG-NER)

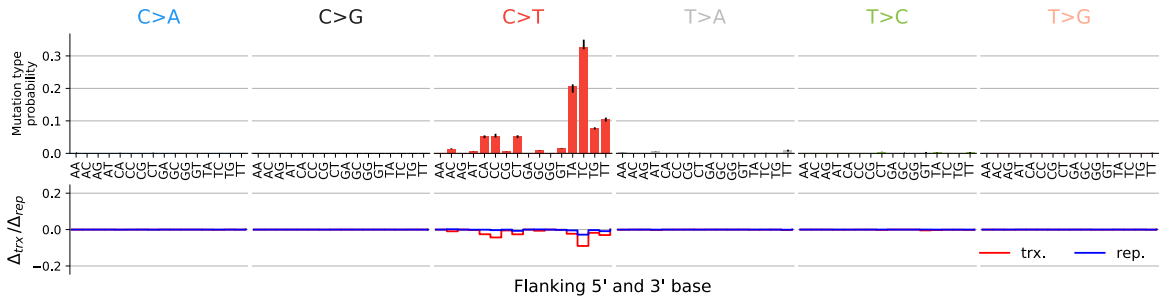


Fig. C.21 **TS05: Single base substitution spectrum.**

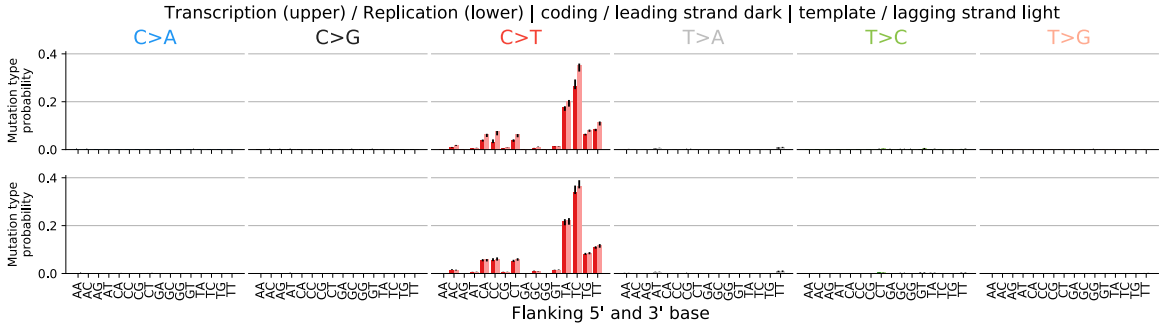


Fig. C.22 **TS05: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

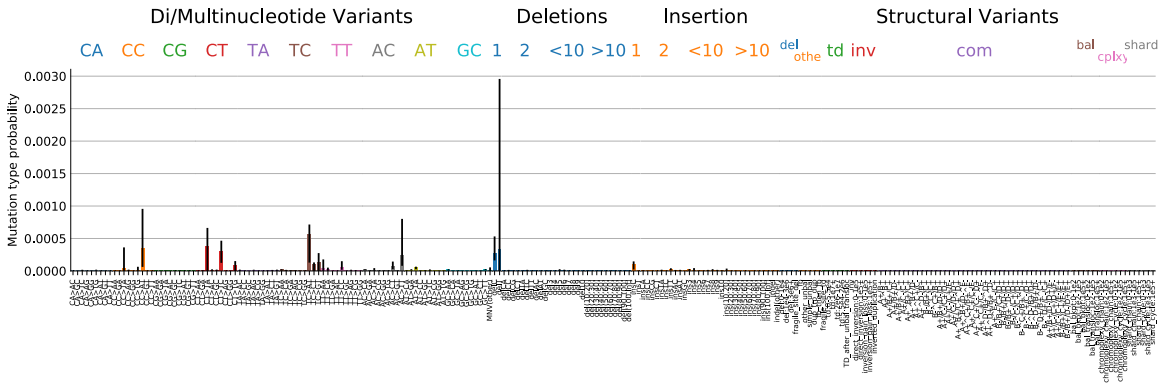


Fig. C.23 **TS05: Spectrum other mutation types.**

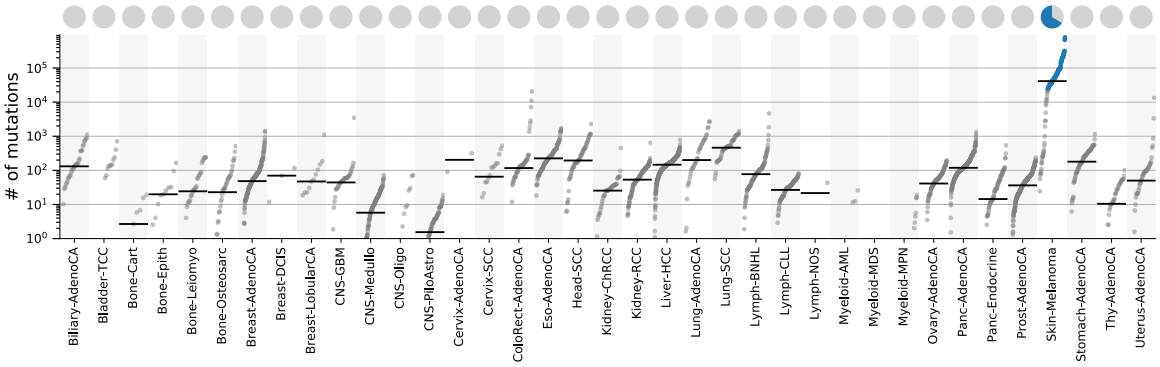


Fig. C.24 **TS05: Signature activity in different cancer types.**

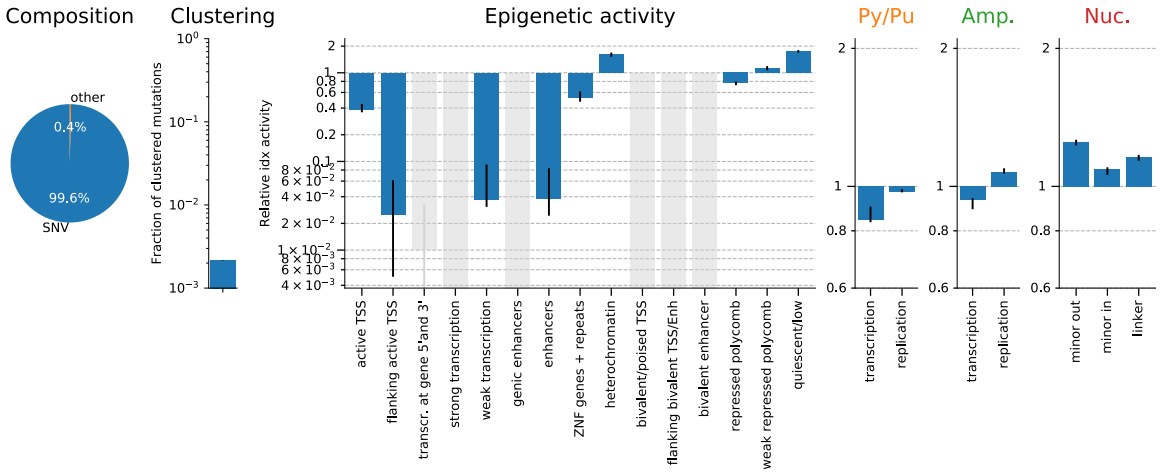


Fig. C.25 **TS05: Signature specific tensor coefficients.**

C.6 TS06-Y[C>T]N (UV/GG+TC-NER)

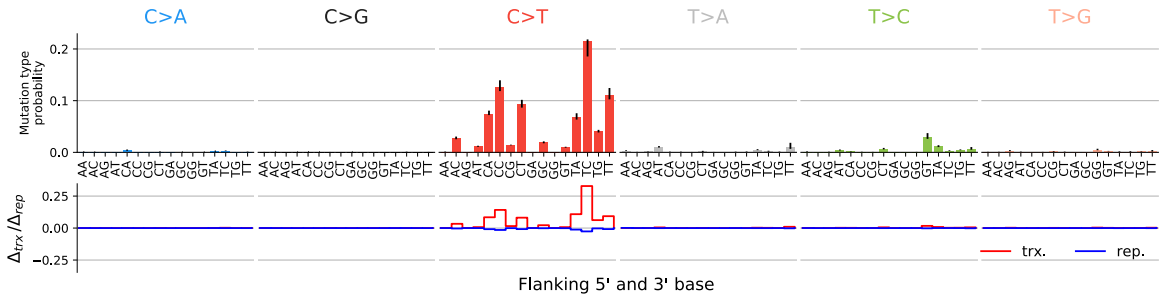


Fig. C.26 **TS06: Single base substitution spectrum.**

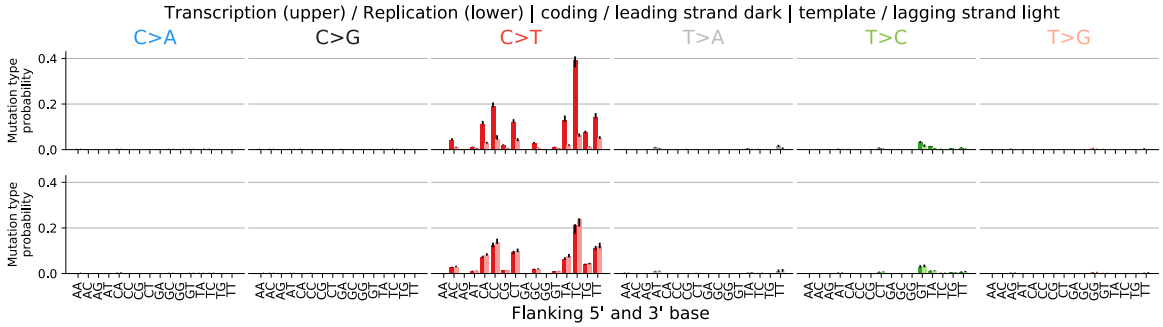


Fig. C.27 **TS06: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

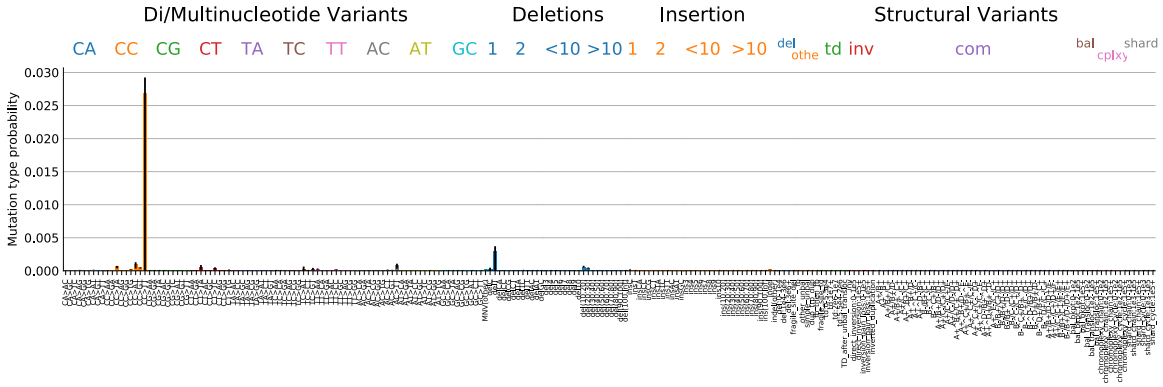


Fig. C.28 **TS06: Spectrum other mutation types.**

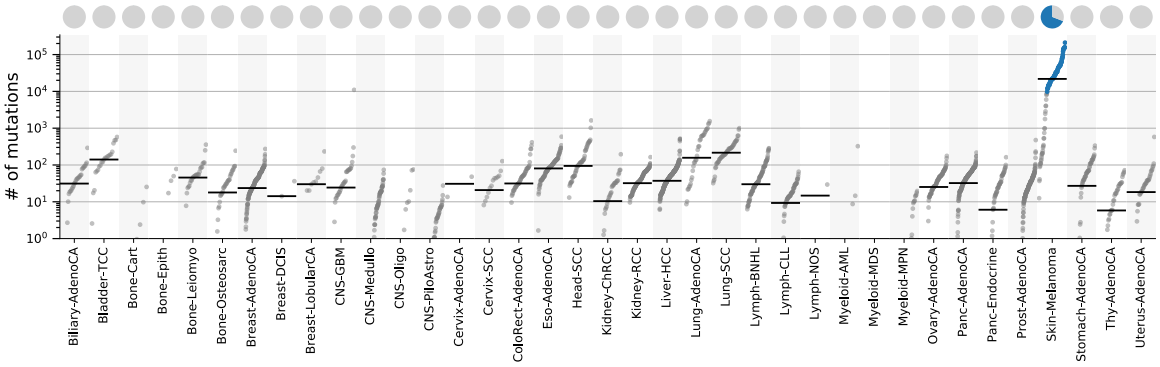


Fig. C.29 **TS06: Signature activity in different cancer types.**

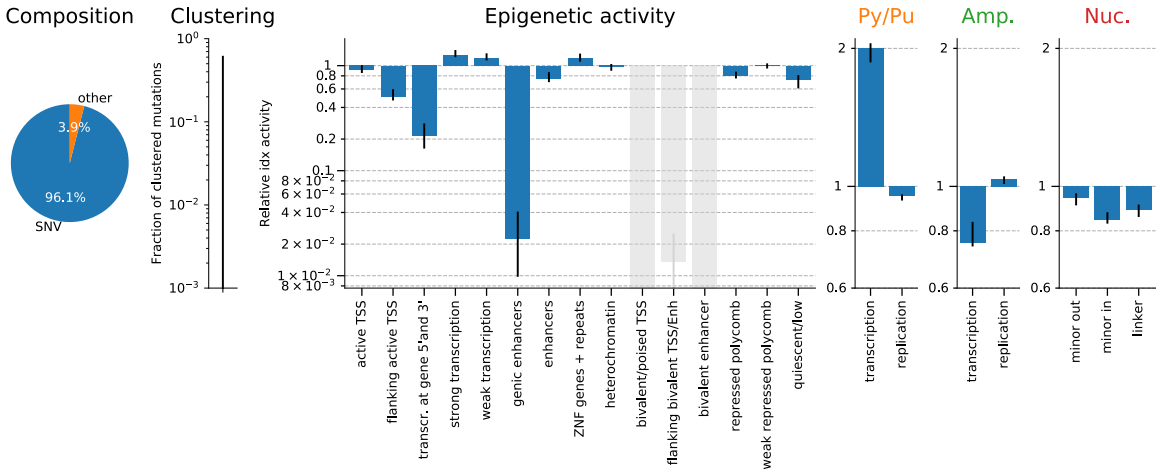


Fig. C.30 **TS06: Signature specific tensor coefficients.**

C.7 TS07-N[T>C]N (unknown)

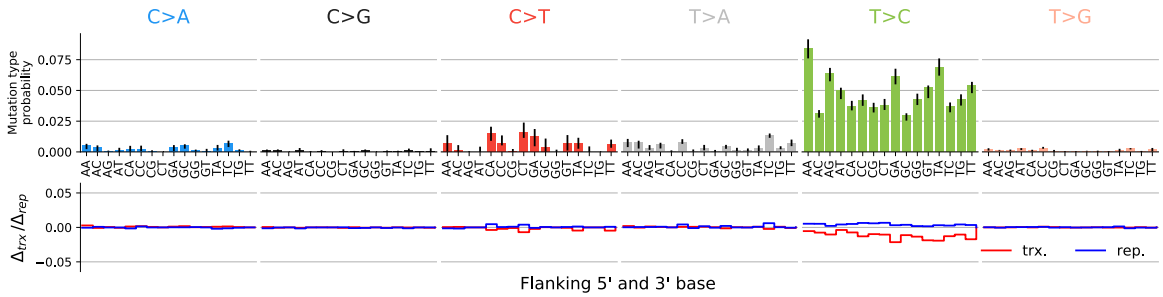


Fig. C.31 **TS07: Single base substitution spectrum.**

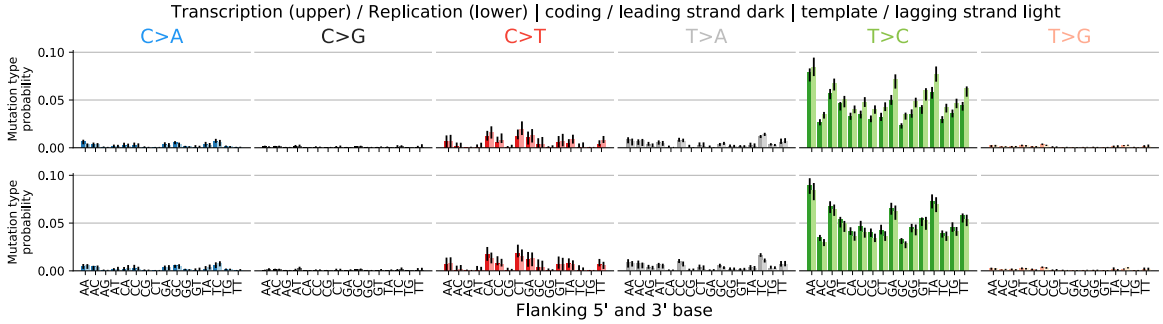


Fig. C.32 **TS07: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

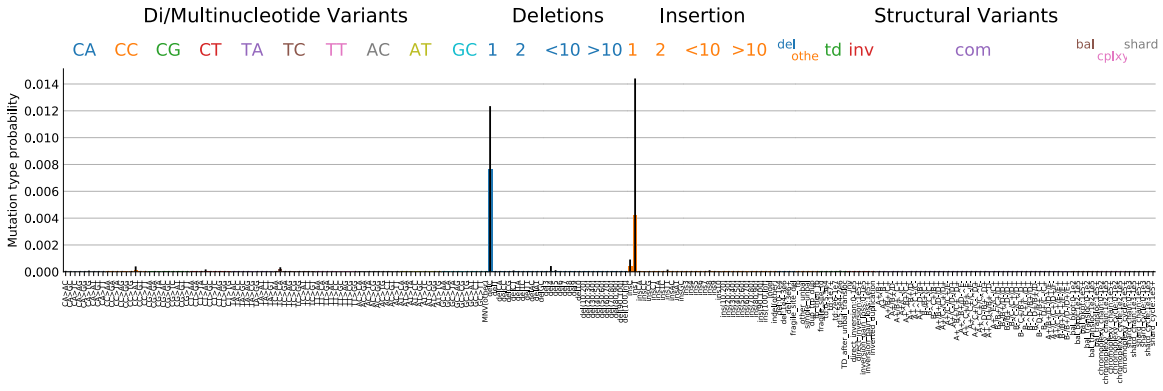


Fig. C.33 **TS07: Spectrum other mutation types.**

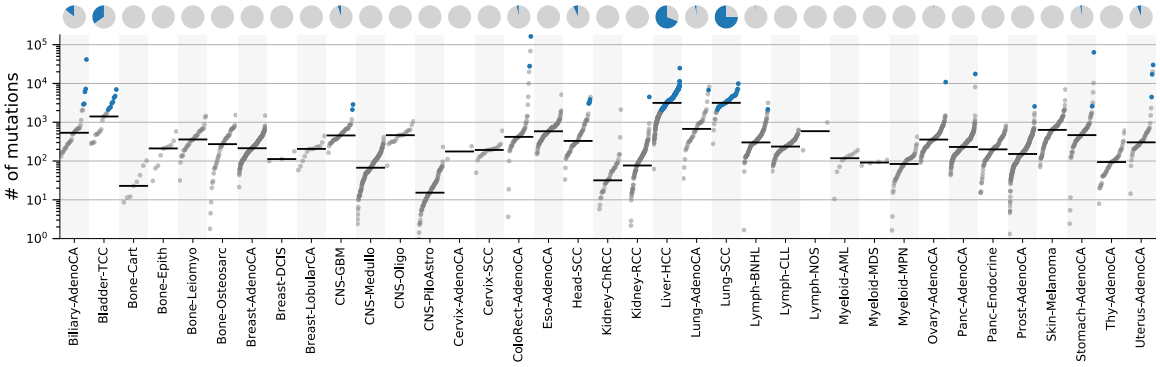


Fig. C.34 **TS07: Signature activity in different cancer types.**

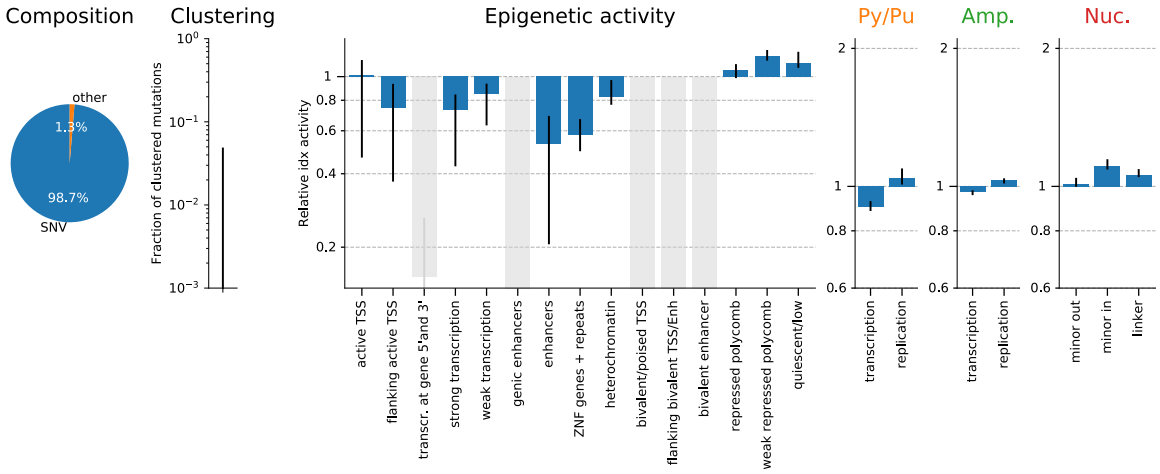


Fig. C.35 **TS07: Signature specific tensor coefficients.**

C.8 TS08-A[T>C]W (unknown/TAM)

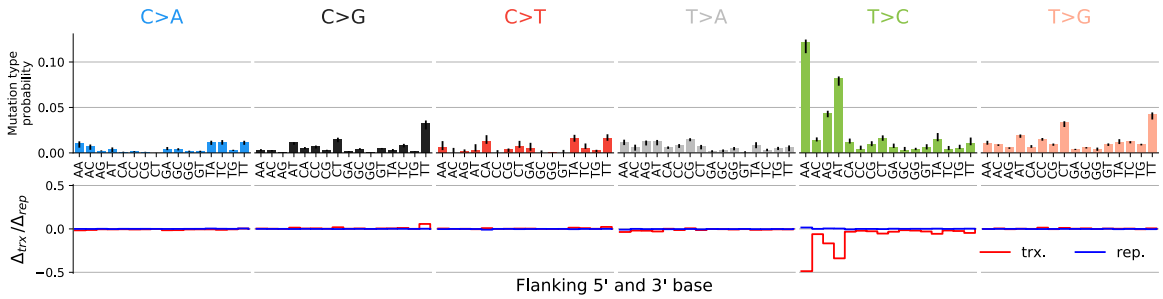


Fig. C.36 **TS08: Single base substitution spectrum.**

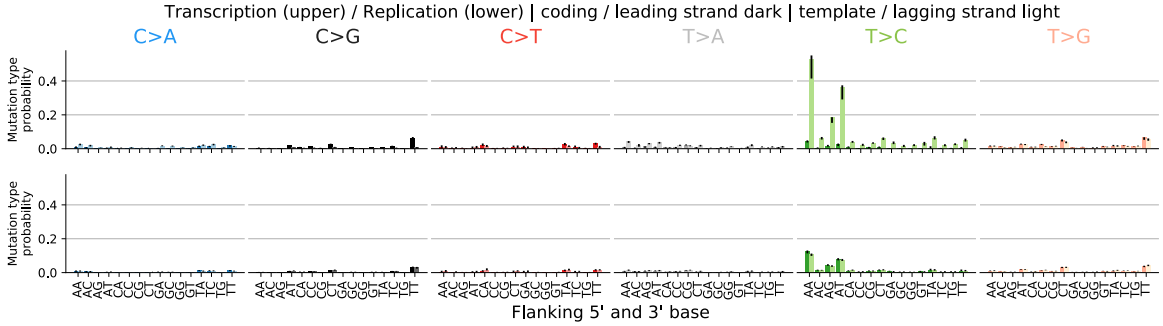


Fig. C.37 **TS08: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

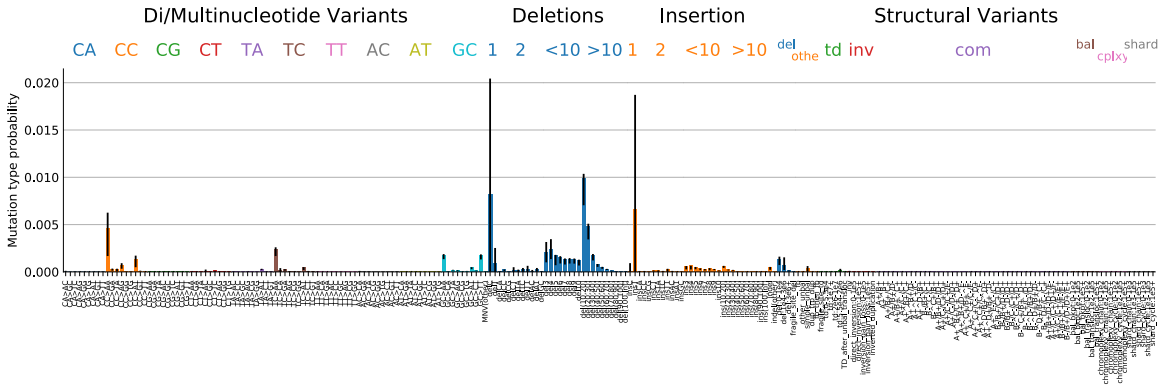


Fig. C.38 **TS08: Spectrum other mutation types.**

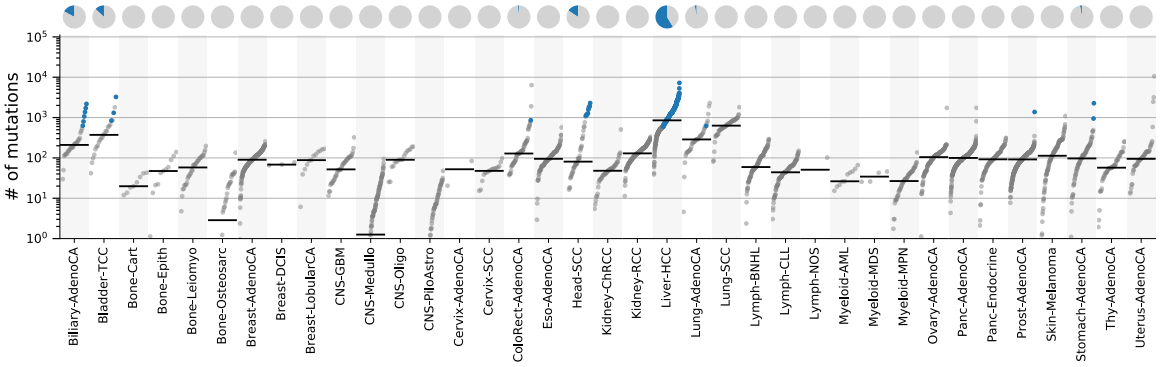


Fig. C.39 **TS08: Signature activity in different cancer types.**

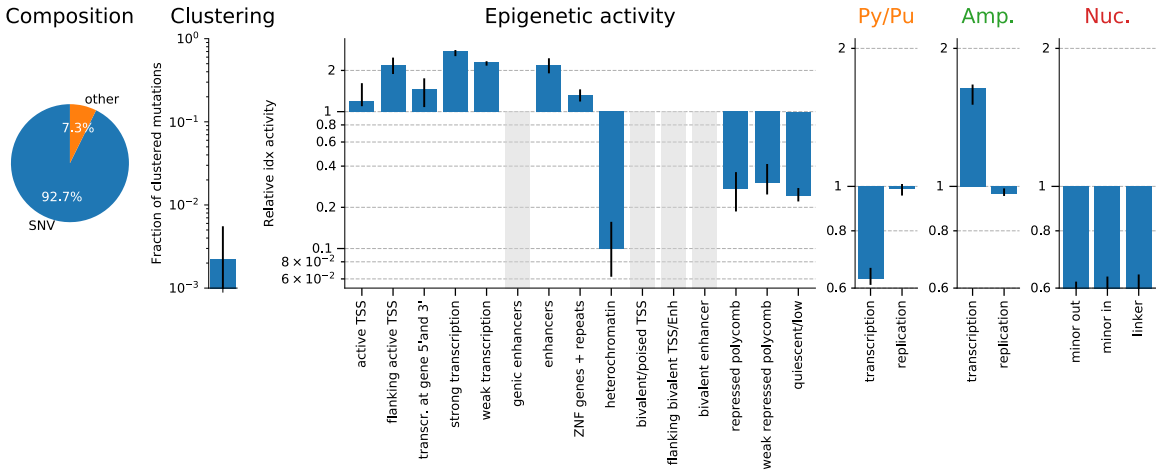


Fig. C.40 **TS08: Signature specific tensor coefficients.**

C.9 TS09-N[T>A]N (PAH/AA)

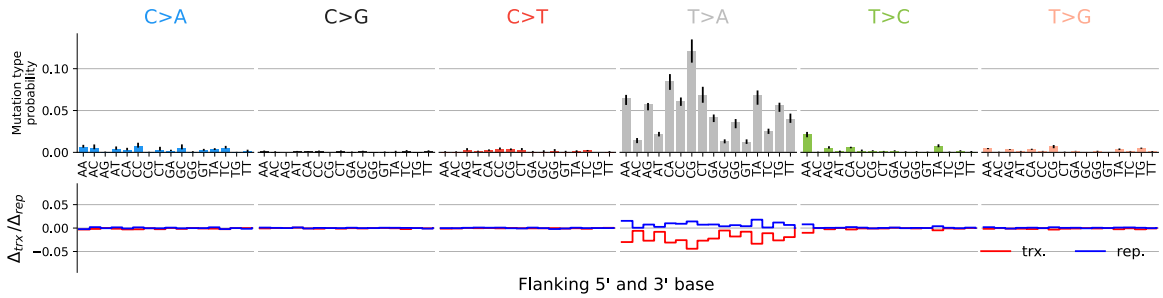


Fig. C.41 **TS09: Single base substitution spectrum.**

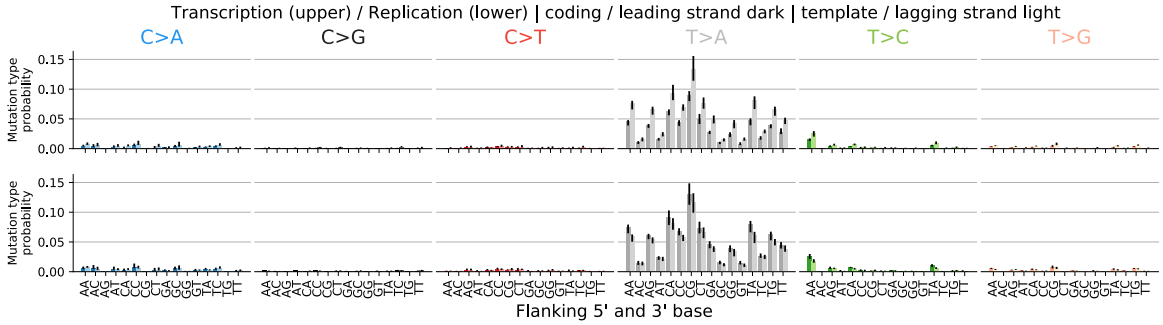


Fig. C.42 **TS09: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

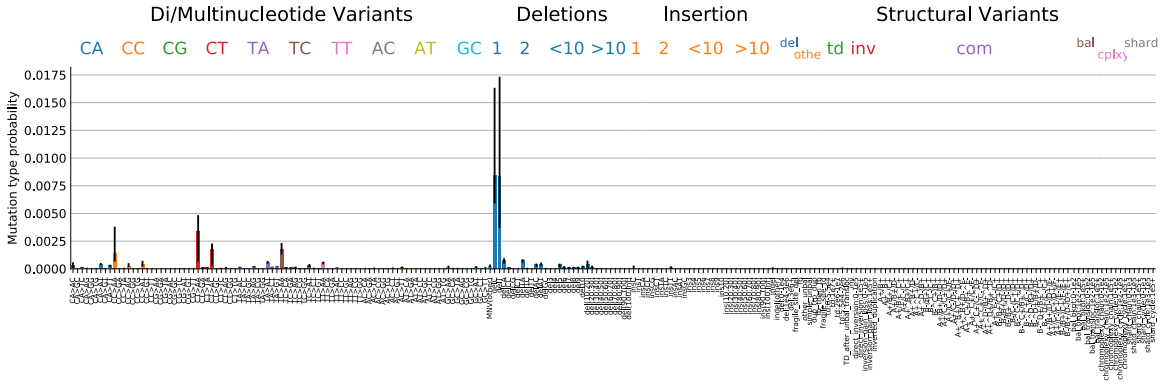


Fig. C.43 **TS09: Spectrum other mutation types.**

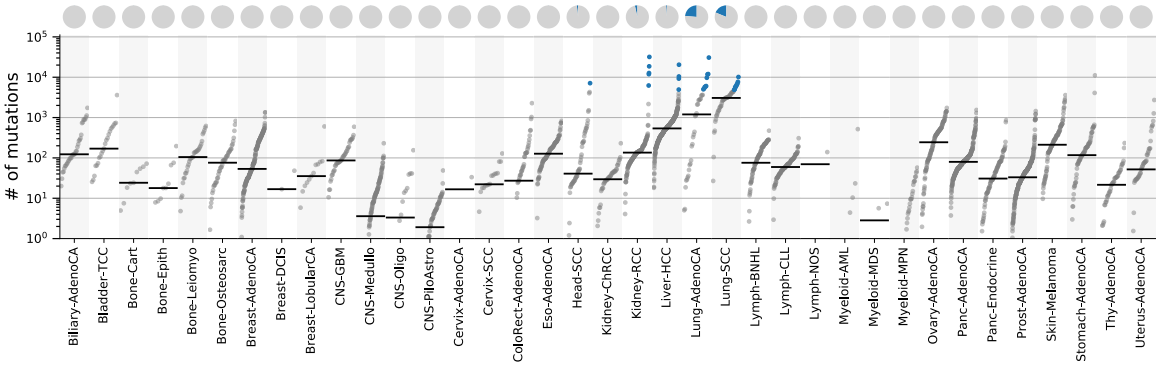


Fig. C.44 **TS09: Signature activity in different cancer types.**

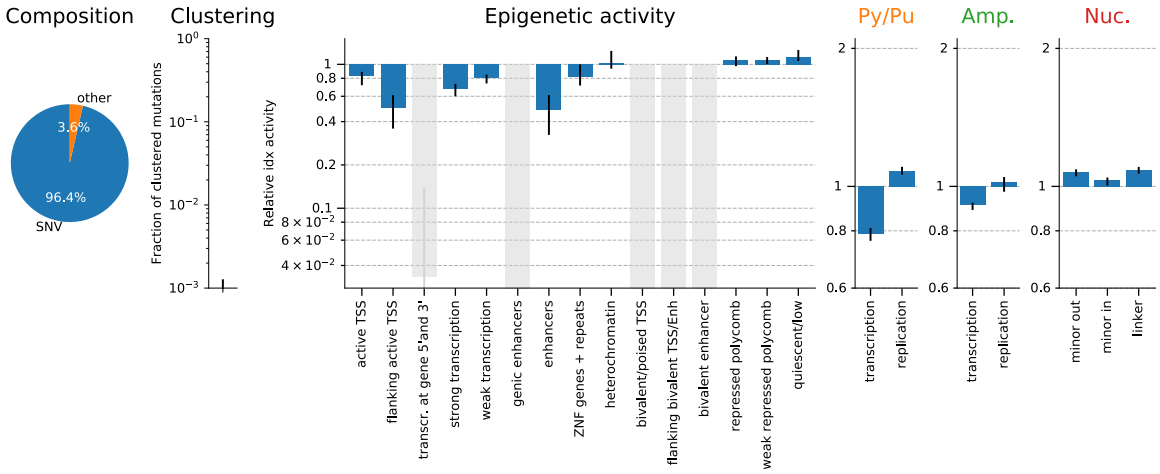


Fig. C.45 **TS09: Signature specific tensor coefficients.**

C.10 TS10-N[C>A]N (PAH/B[a]P)

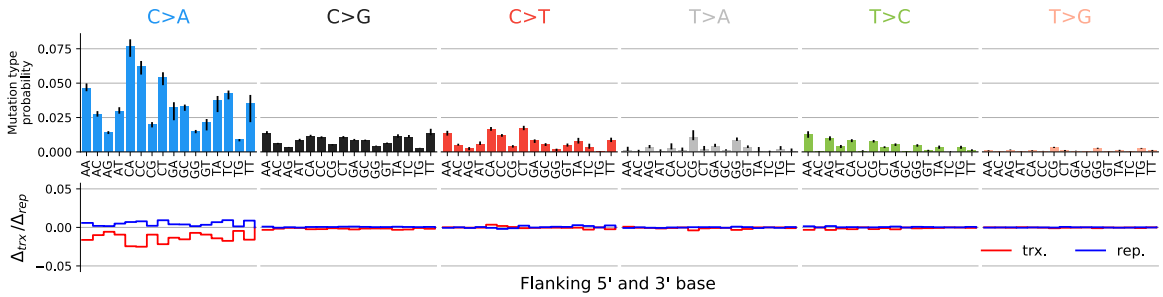


Fig. C.46 **TS10: Single base substitution spectrum.**

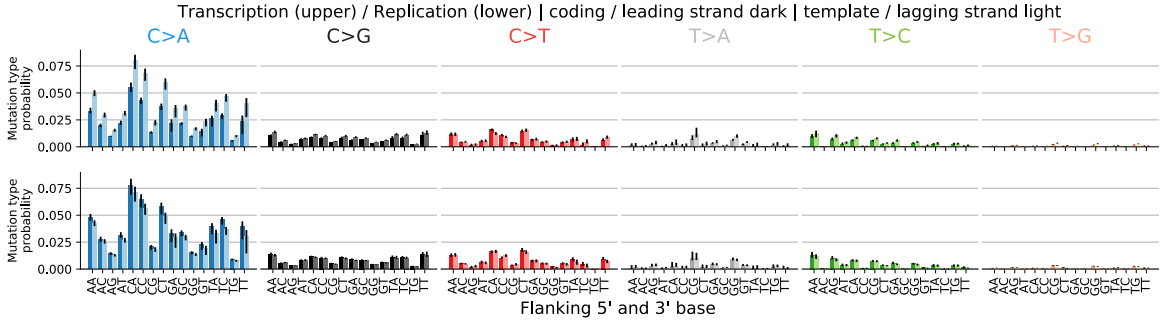


Fig. C.47 **TS10: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

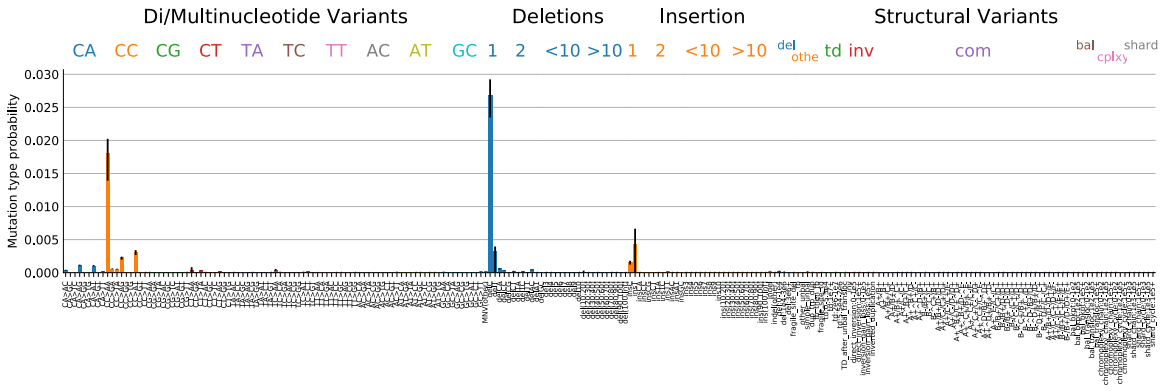


Fig. C.48 **TS10: Spectrum other mutation types.**

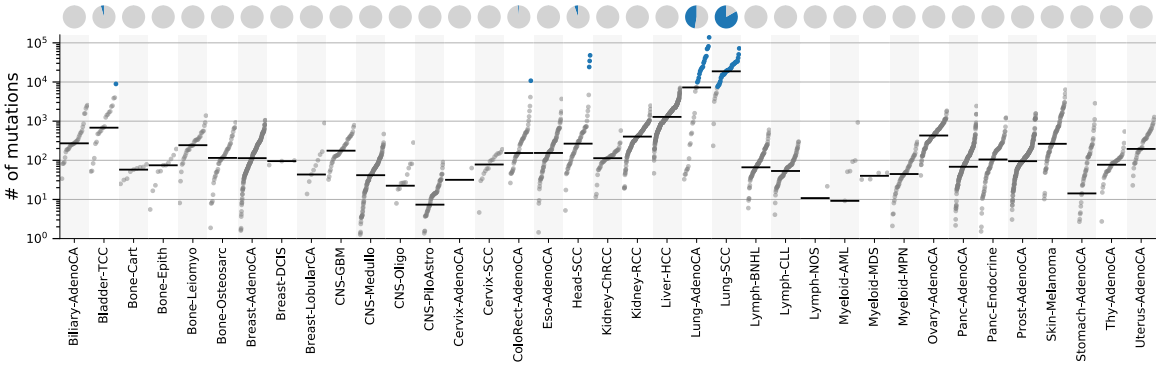


Fig. C.49 **TS10: Signature activity in different cancer types.**

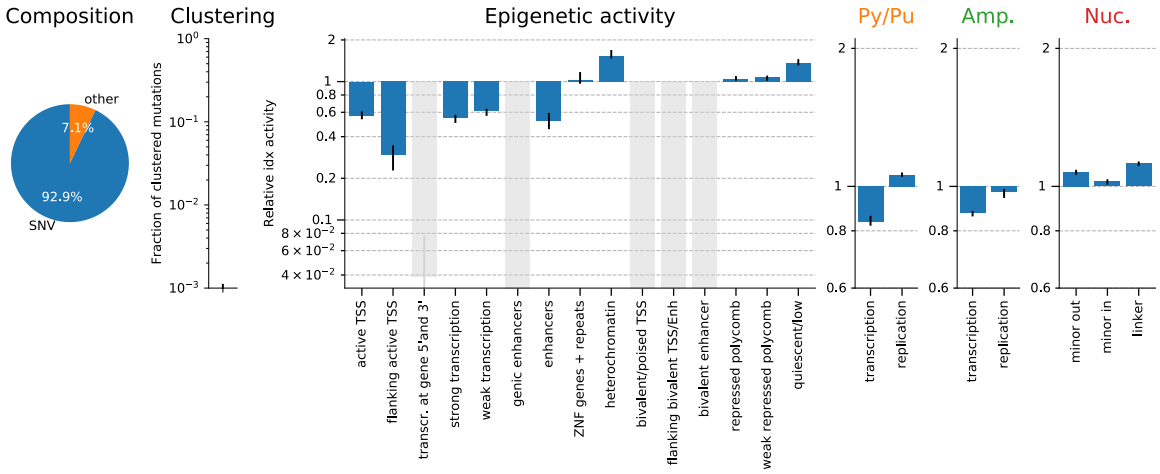


Fig. C.50 **TS10: Signature specific tensor coefficients.**

C.11 TS11-T[C>D]W;SV (APOBEC)

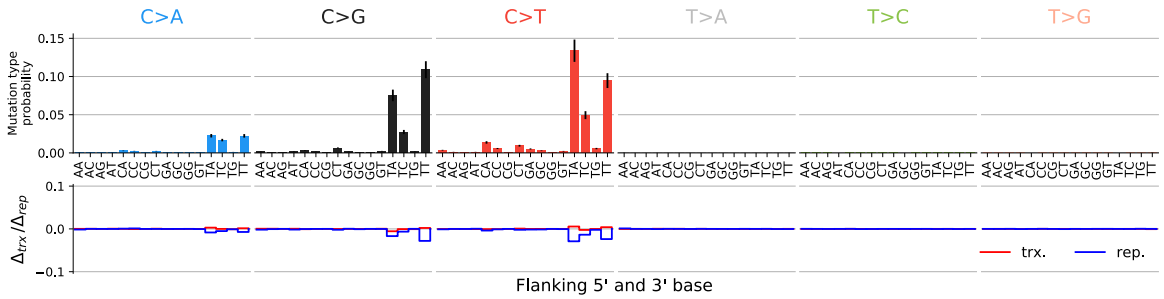


Fig. C.51 **TS11: Single base substitution spectrum.**

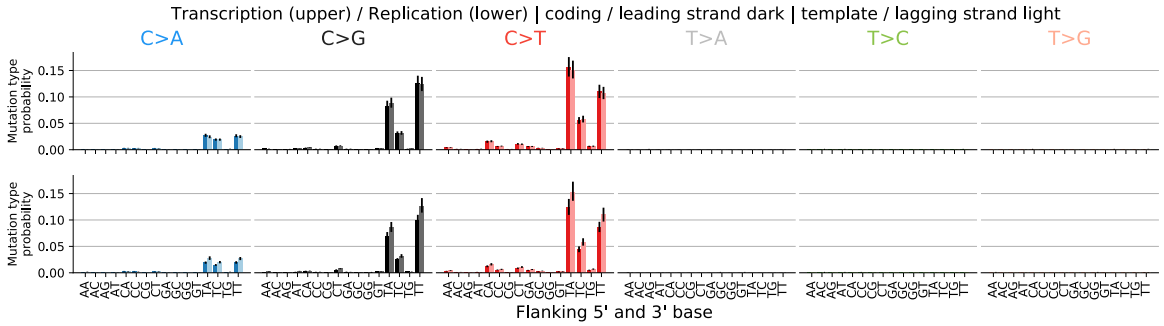


Fig. C.52 **TS11: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

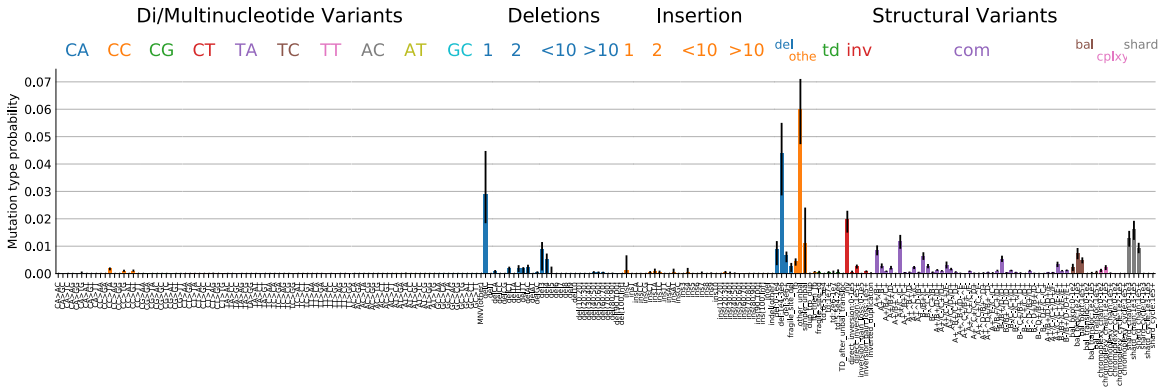


Fig. C.53 **TS11: Spectrum other mutation types.**

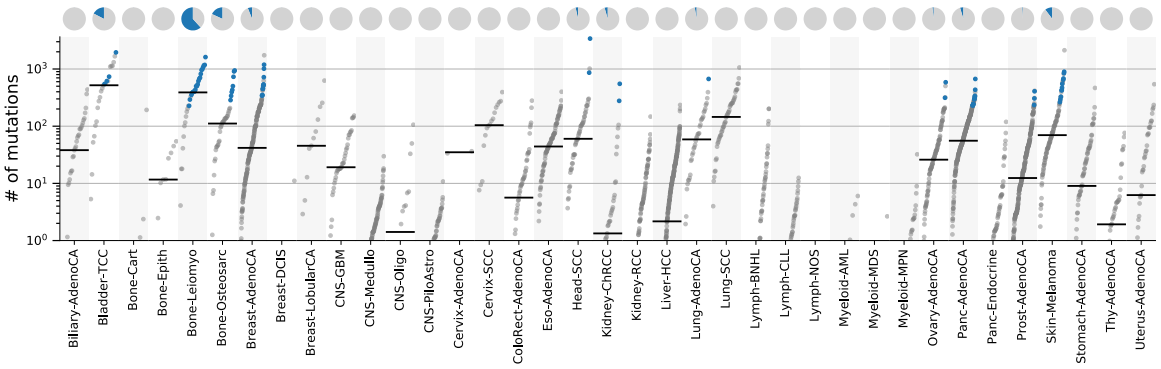


Fig. C.54 **TS11: Signature activity in different cancer types.**

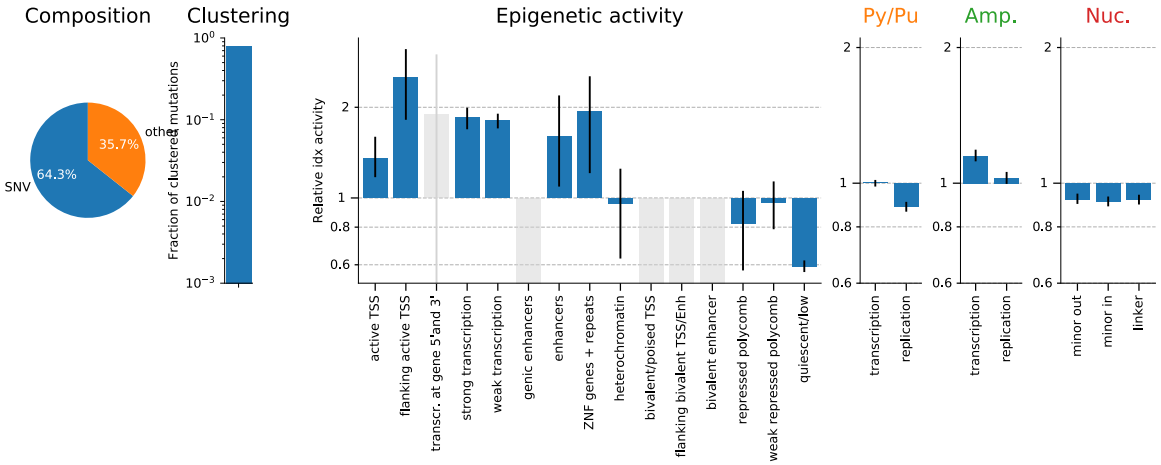


Fig. C.55 **TS11: Signature specific tensor coefficients.**

C.12 TS12-T[C>D]W (APOBEC)

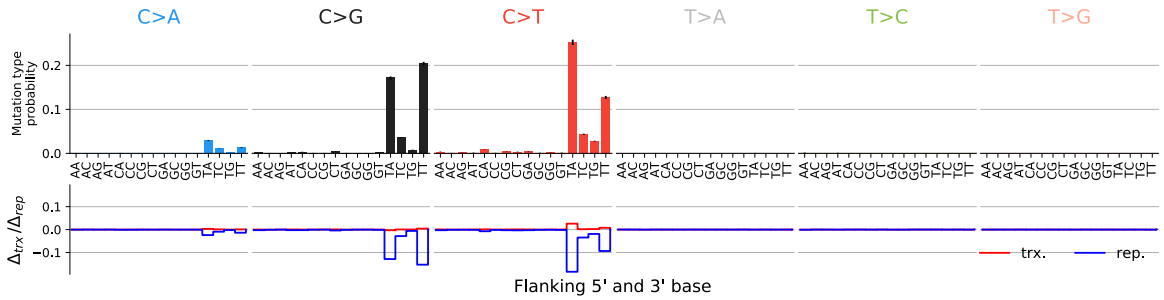


Fig. C.56 **TS12: Single base substitution spectrum.**

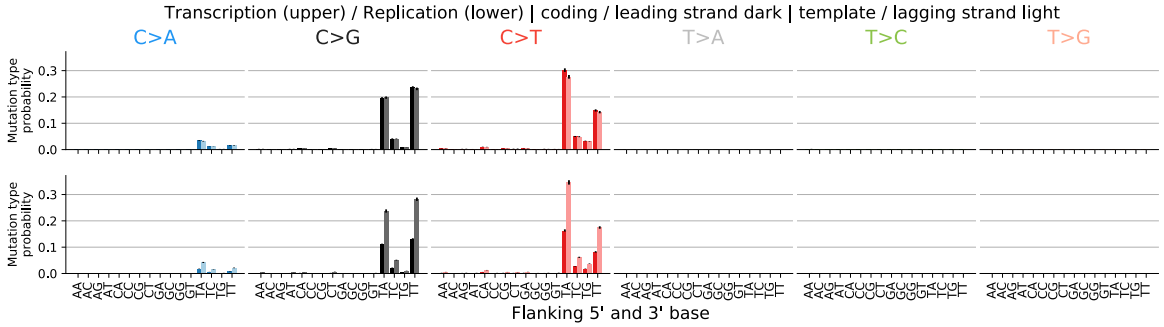


Fig. C.57 **TS12: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

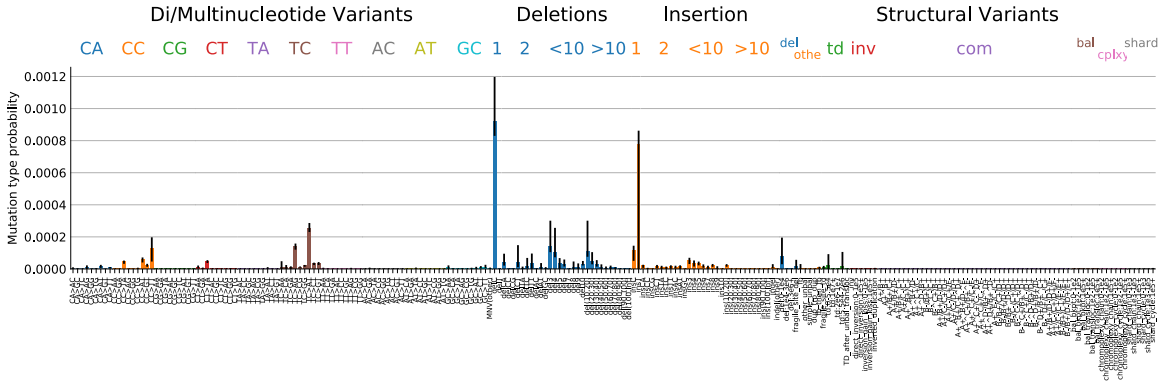


Fig. C.58 **TS12: Spectrum other mutation types.**

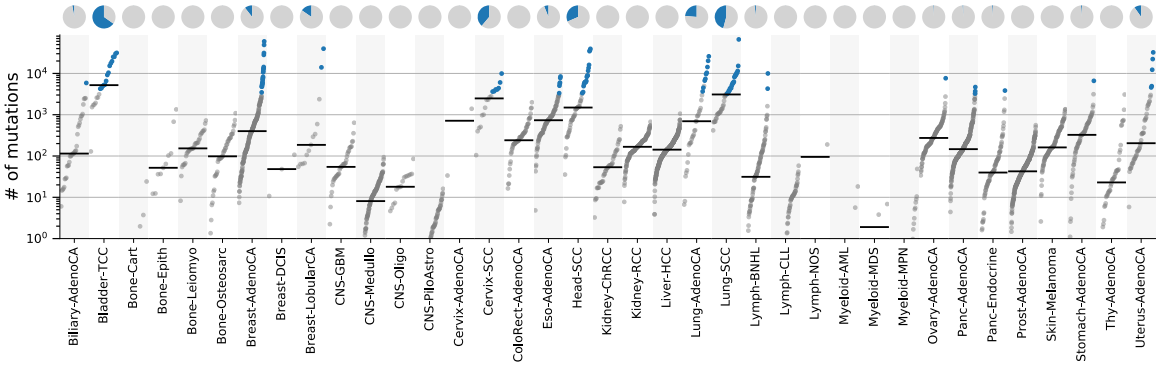


Fig. C.59 **TS12: Signature activity in different cancer types.**

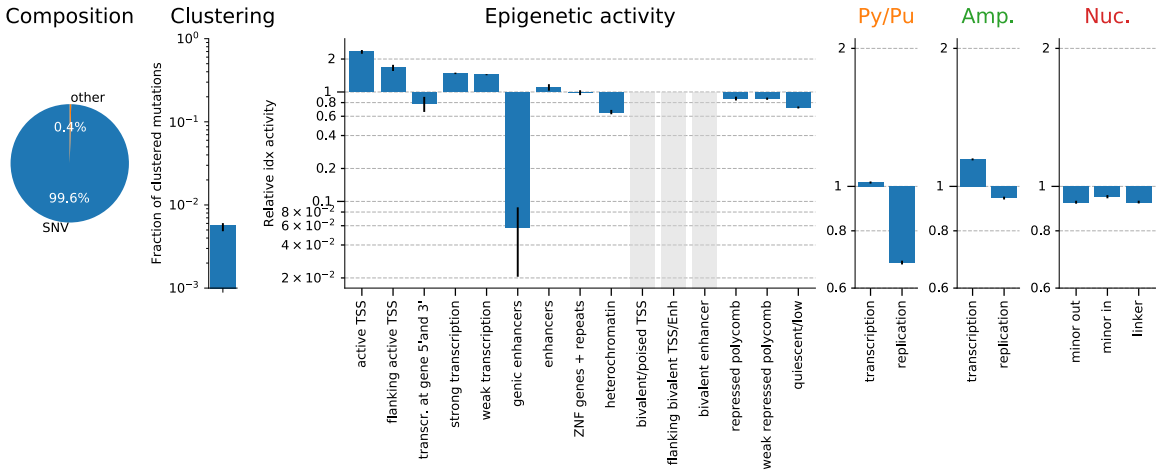


Fig. C.60 **TS12: Signature specific tensor coefficients.**

C.13 TS13-N[C>K]H (AID/SHM)

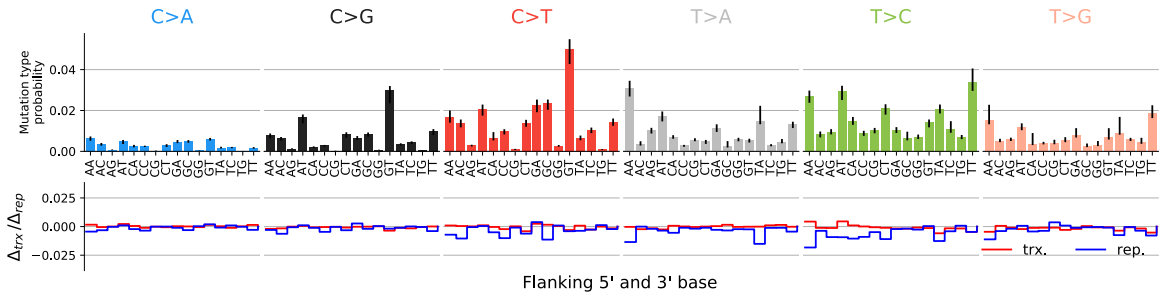


Fig. C.61 **TS13: Single base substitution spectrum.**

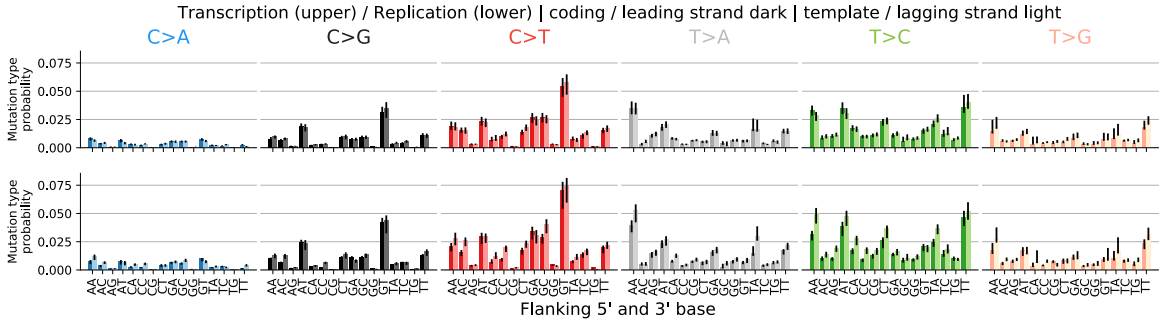


Fig. C.62 **TS13: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

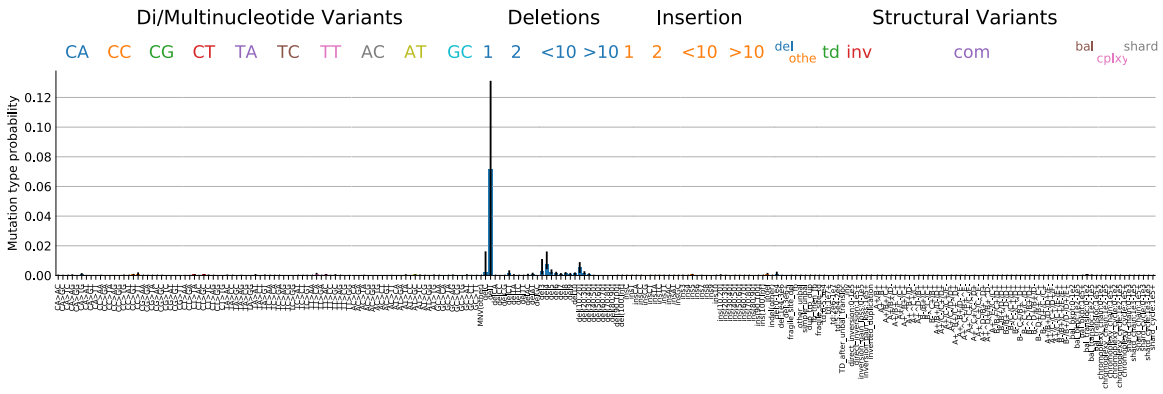


Fig. C.63 **TS13: Spectrum other mutation types.**

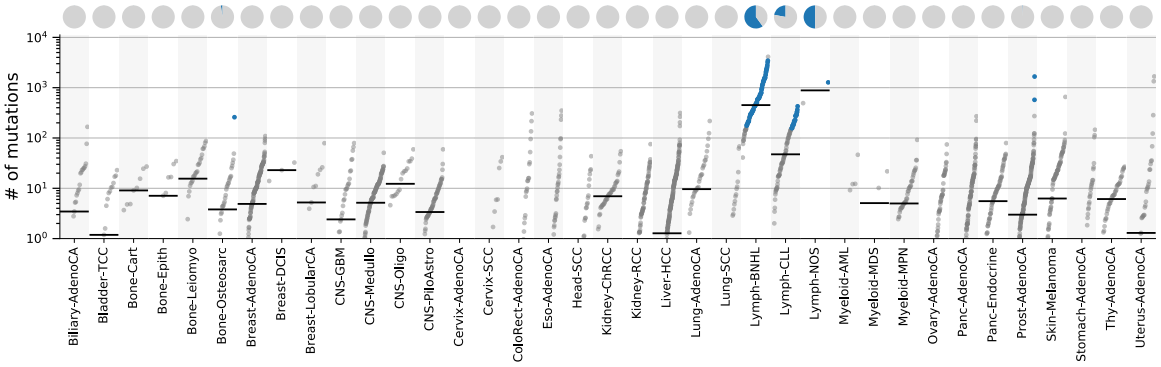


Fig. C.64 **TS13: Signature activity in different cancer types.**

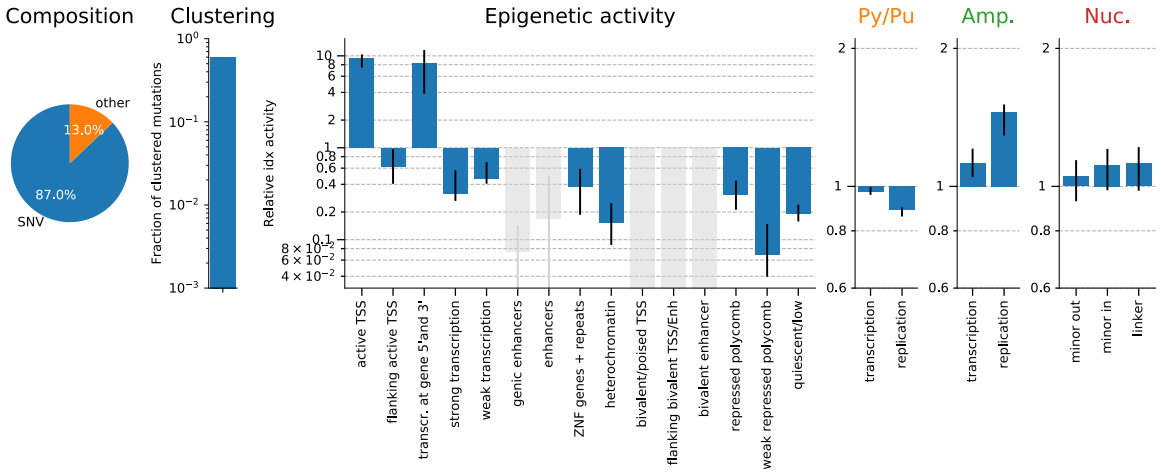


Fig. C.65 **TS13: Signature specific tensor coefficients.**

C.14 TS14-W[T>V]W (POLH)

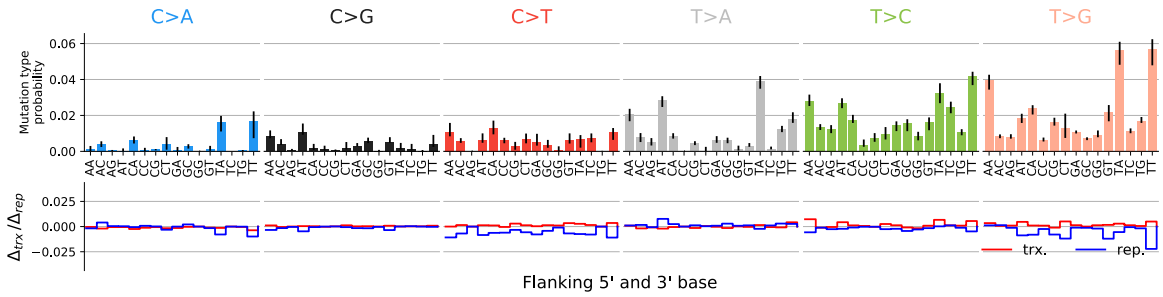


Fig. C.66 **TS14: Single base substitution spectrum.**

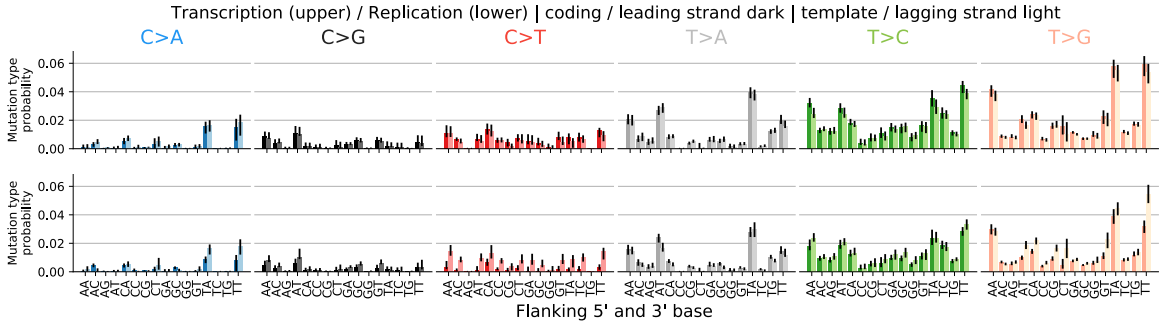


Fig. C.67 **TS14: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

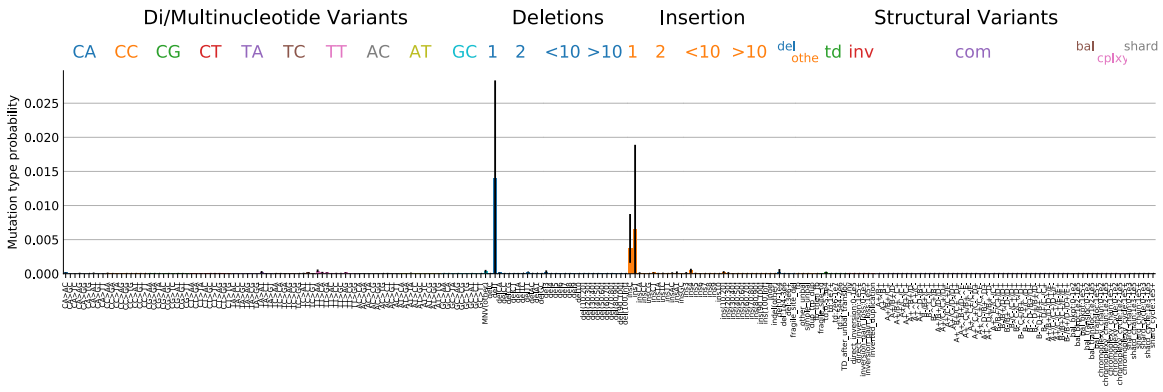


Fig. C.68 **TS14: Spectrum other mutation types.**

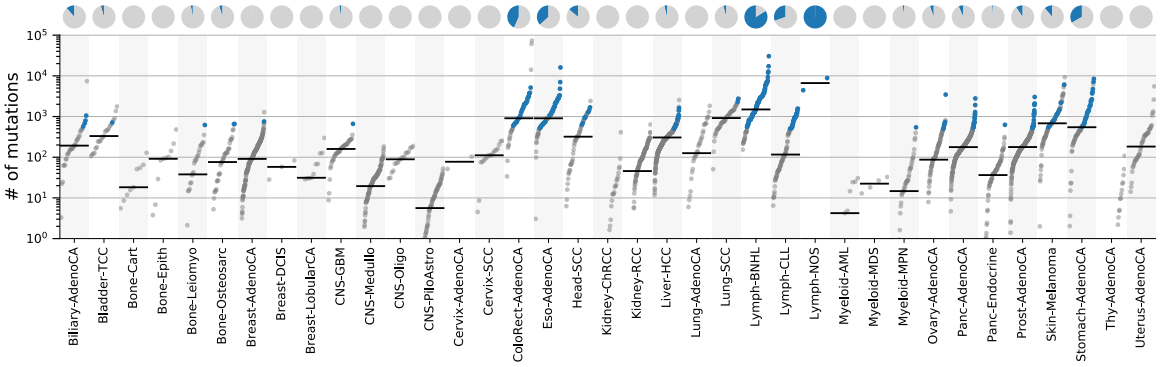


Fig. C.69 **TS14: Signature activity in different cancer types.**

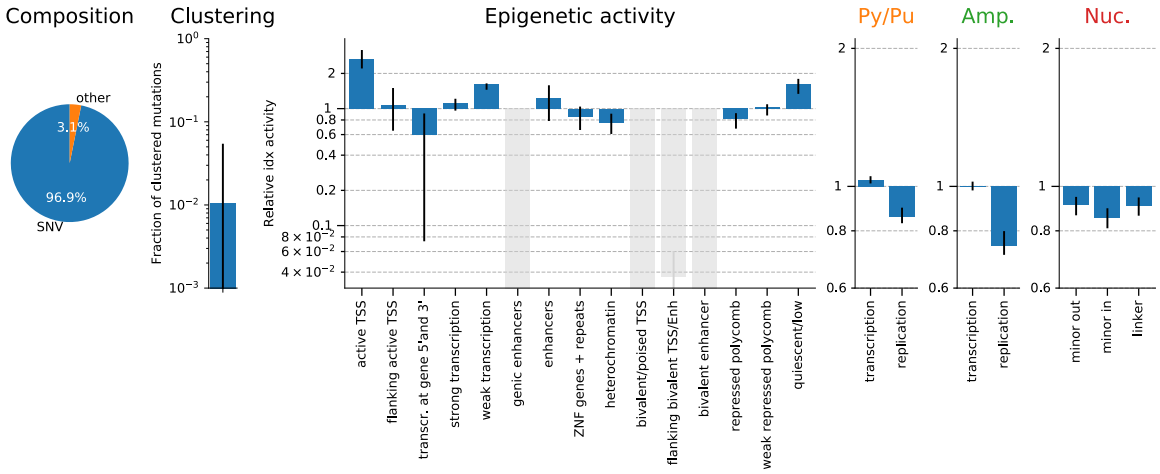


Fig. C.70 **TS14: Signature specific tensor coefficients.**

C.15 TS15-G[C>T]N;ID (MMRD)

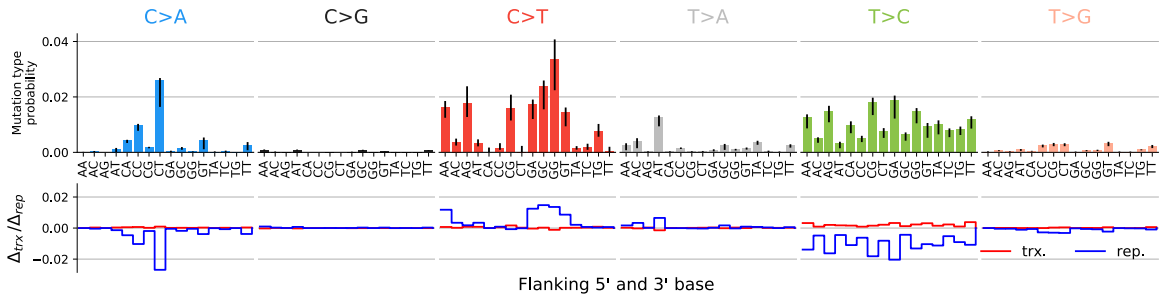


Fig. C.71 **TS15: Single base substitution spectrum.**

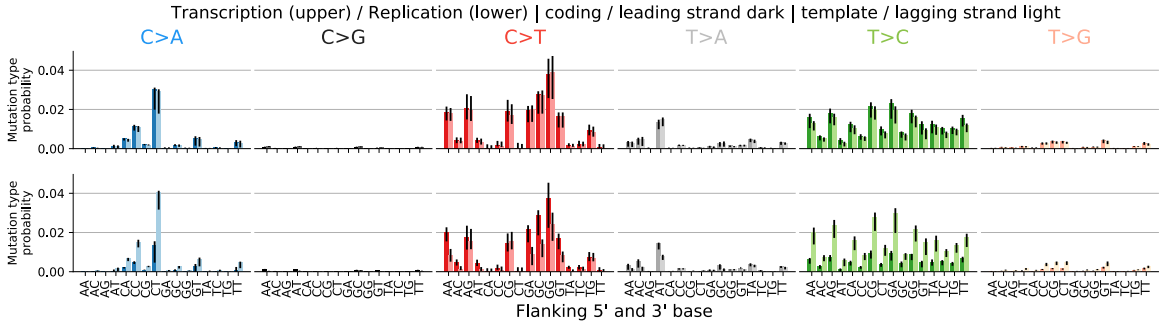


Fig. C.72 **TS15: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

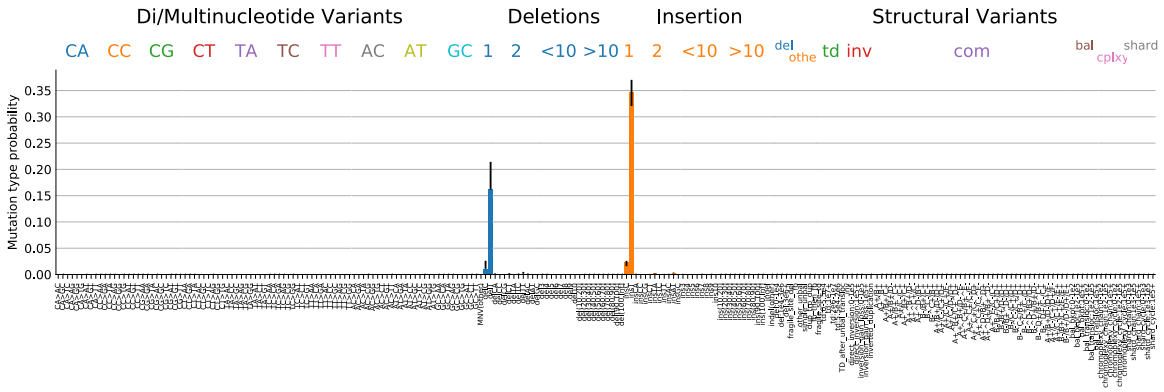


Fig. C.73 **TS15: Spectrum other mutation types.**

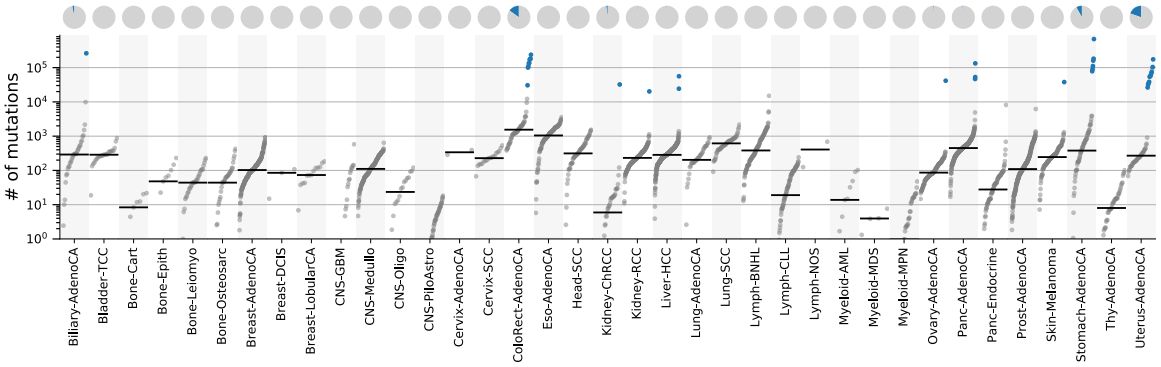


Fig. C.74 **TS15: Signature activity in different cancer types.**

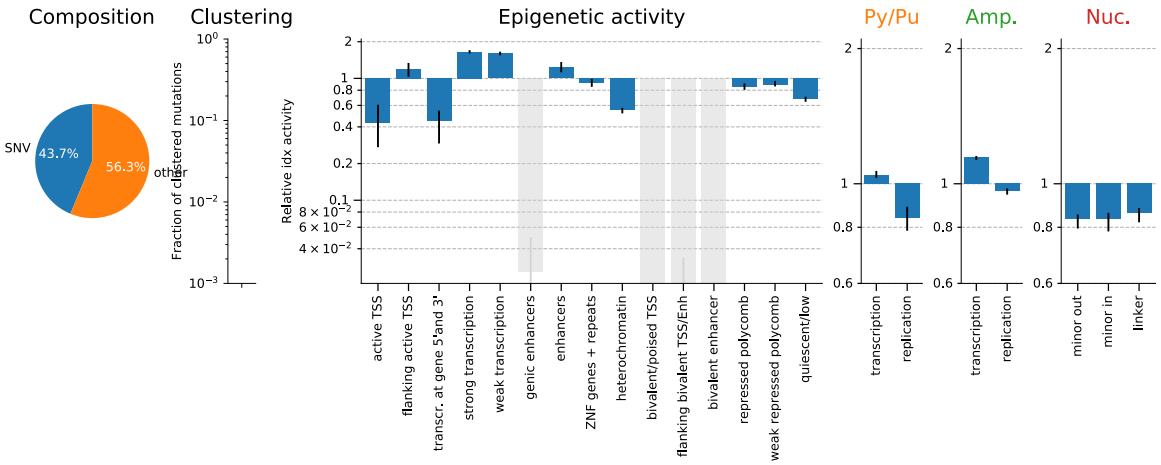


Fig. C.75 **TS15: Signature specific tensor coefficients.**

C.16 TS16-N[C>A]T;ID (MMRD:POLE-exo)

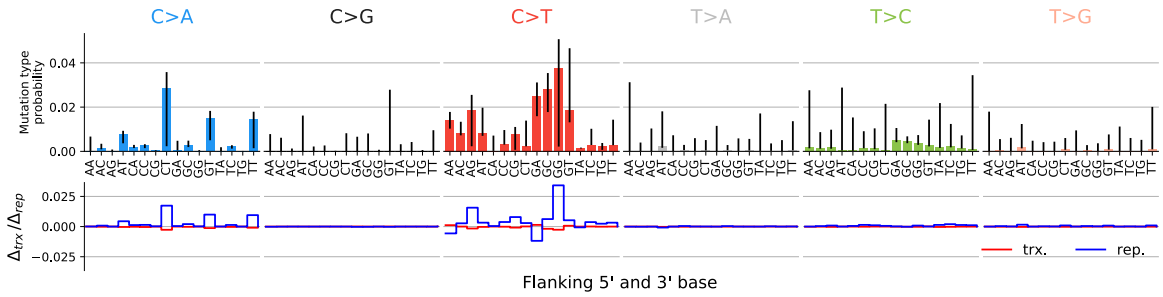


Fig. C.76 **TS16: Single base substitution spectrum.**

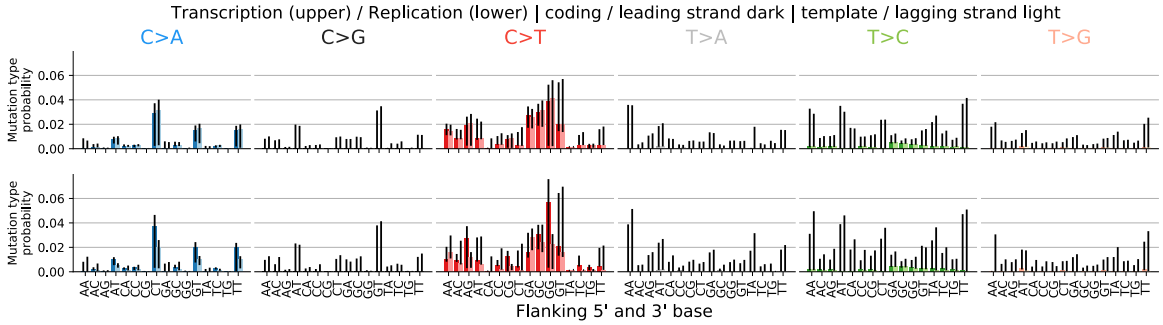


Fig. C.77 **TS16: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

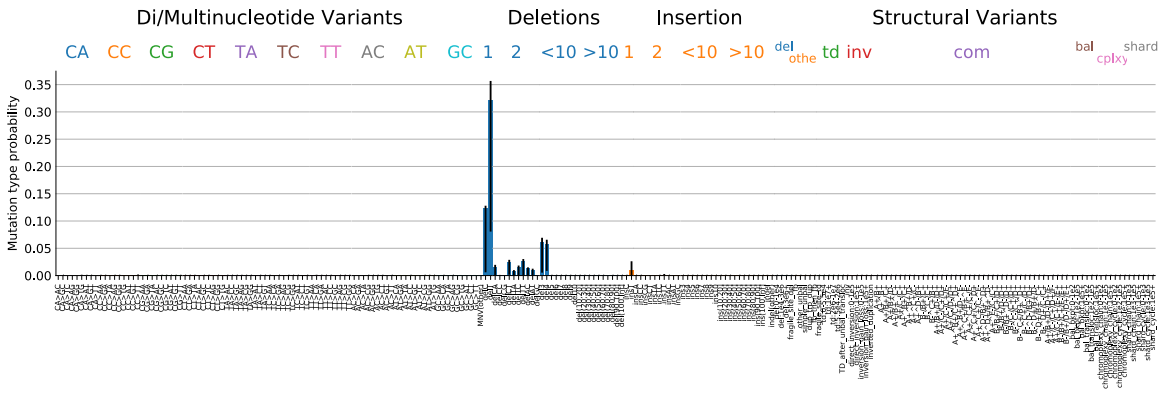


Fig. C.78 **TS16: Spectrum other mutation types.**

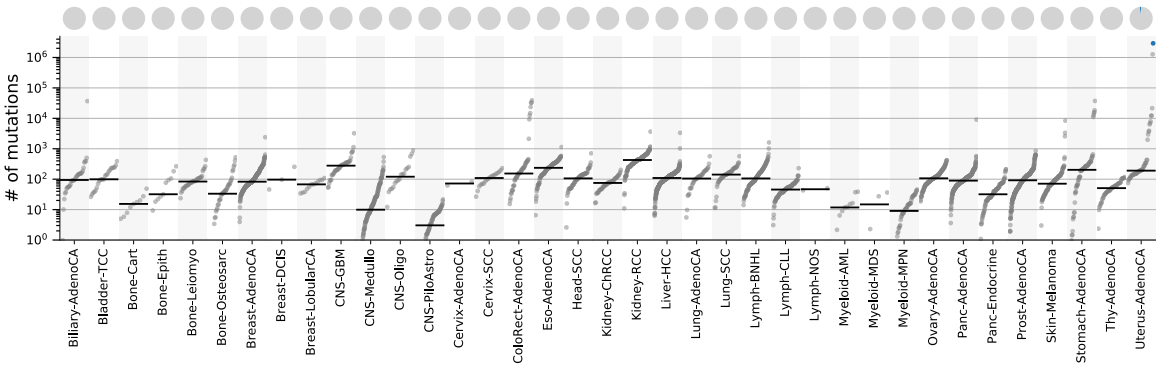


Fig. C.79 **TS16: Signature activity in different cancer types.**

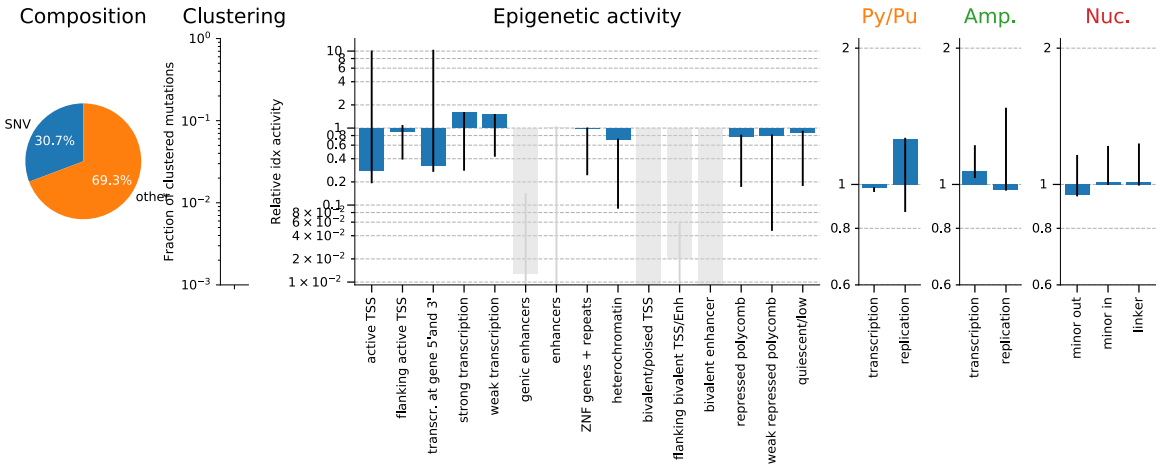


Fig. C.80 **TS16: Signature specific tensor coefficients.**

C.17 TS17-T[C>A]T (POLE-exo)

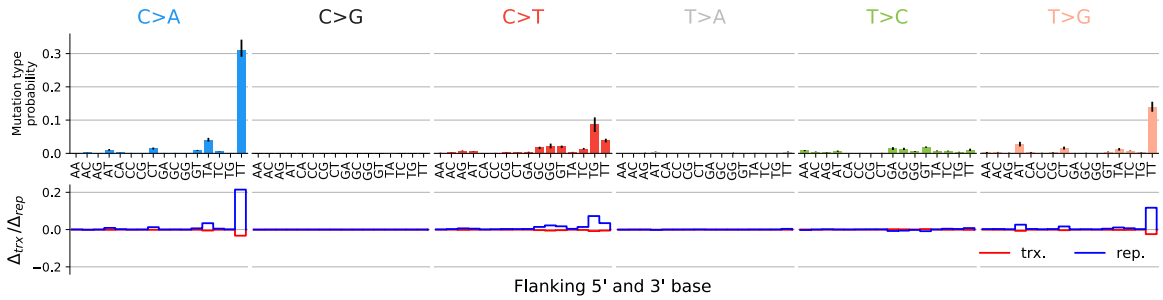


Fig. C.81 **TS17: Single base substitution spectrum.**

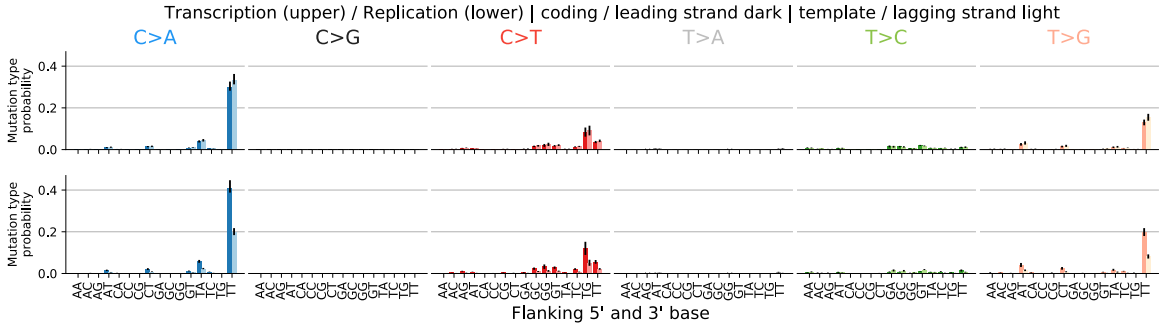


Fig. C.82 **TS17: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

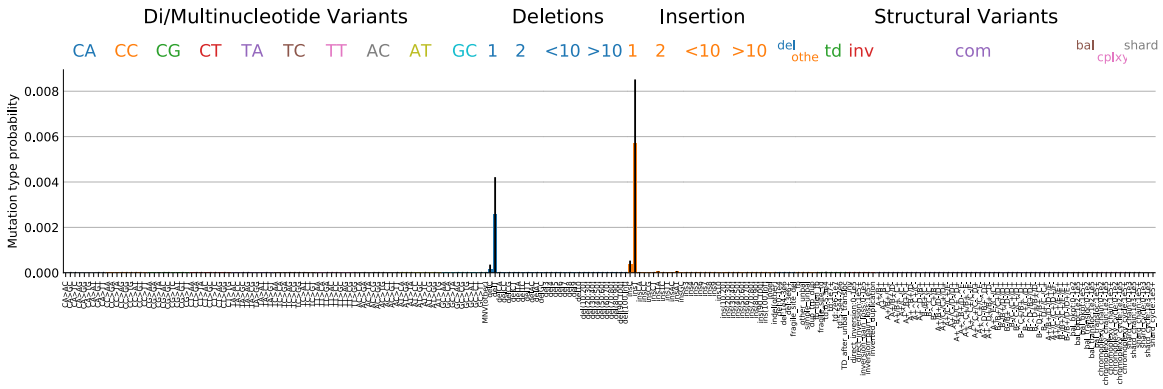


Fig. C.83 **TS17: Spectrum other mutation types.**

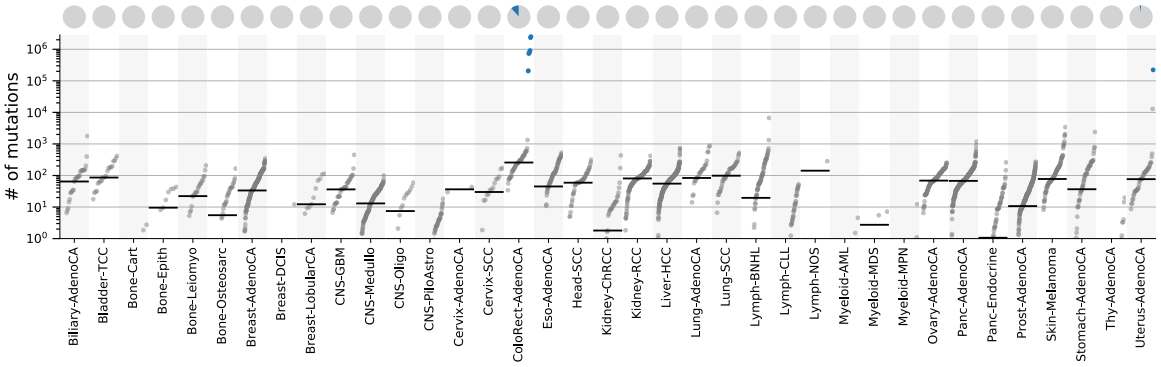


Fig. C.84 **TS17: Signature activity in different cancer types.**

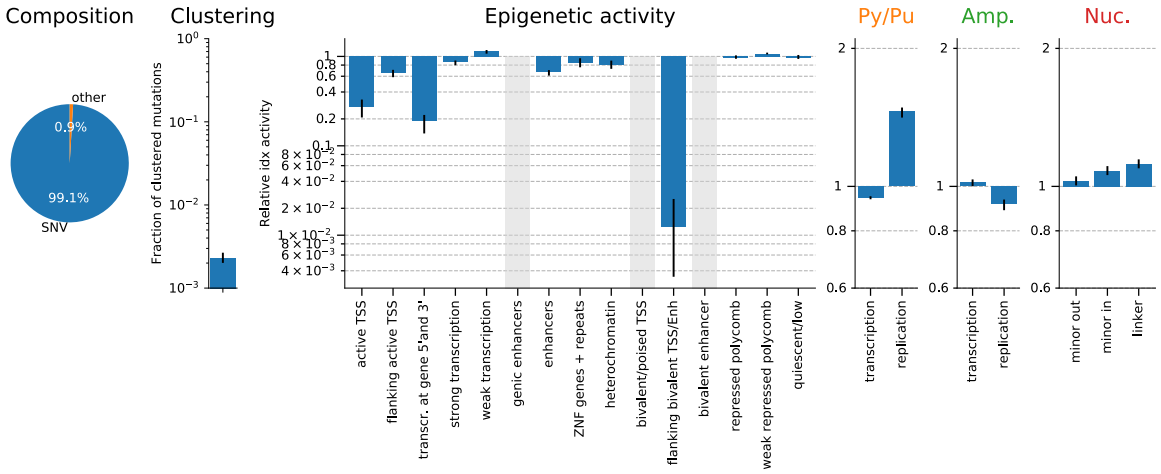


Fig. C.85 **TS17: Signature specific tensor coefficients.**

C.18 TS18-N[C>A]W (BERD/MUTYH)

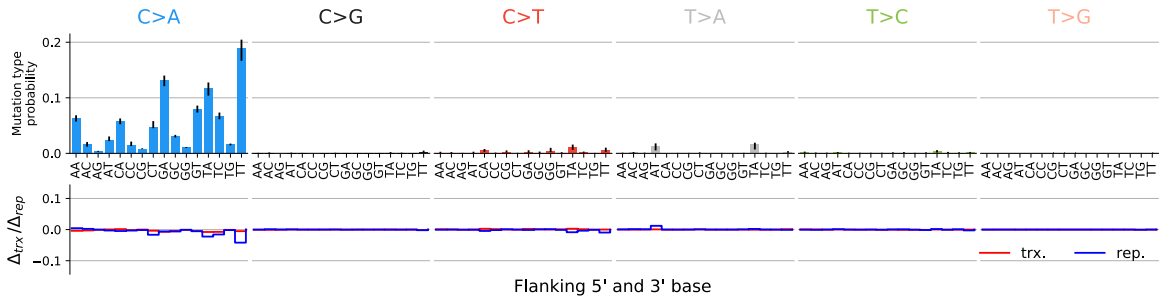


Fig. C.86 **TS18: Single base substitution spectrum.**

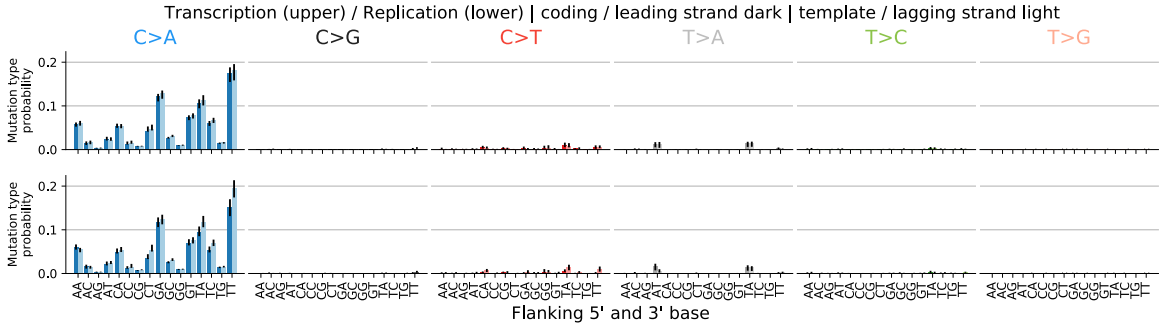


Fig. C.87 **TS18: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

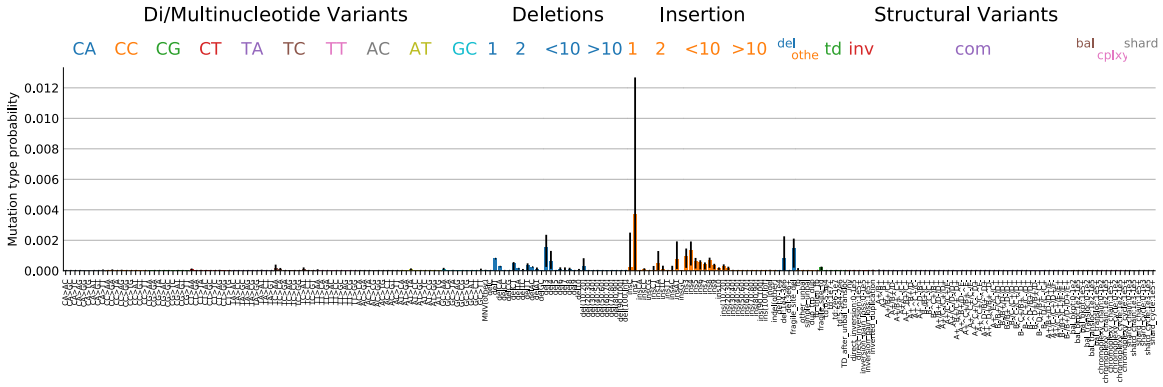


Fig. C.88 **TS18: Spectrum other mutation types.**

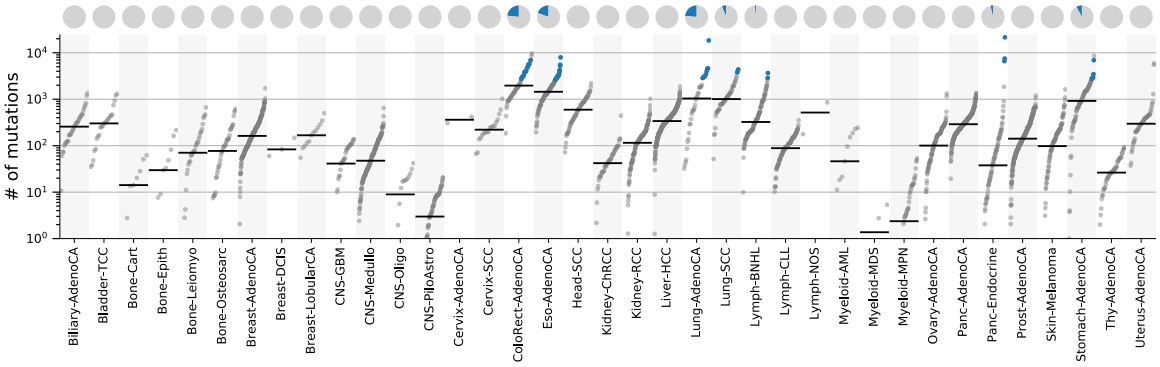


Fig. C.89 **TS18: Signature activity in different cancer types.**

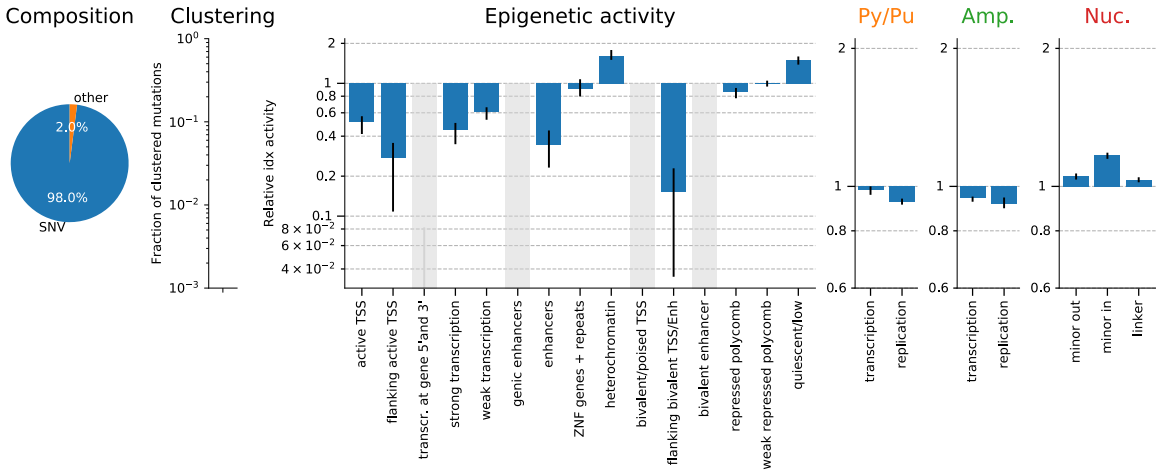


Fig. C.90 **TS18: Signature specific tensor coefficients.**

C.19 TS19-N[N>N]N;SV (HRD/BRCA)

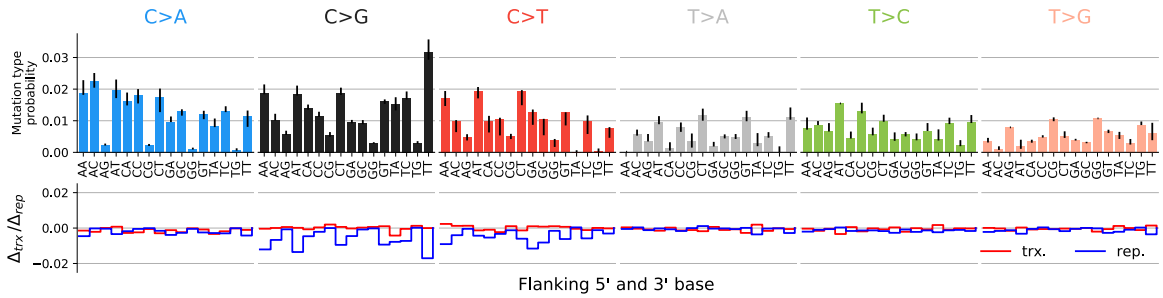


Fig. C.91 **TS19: Single base substitution spectrum.**

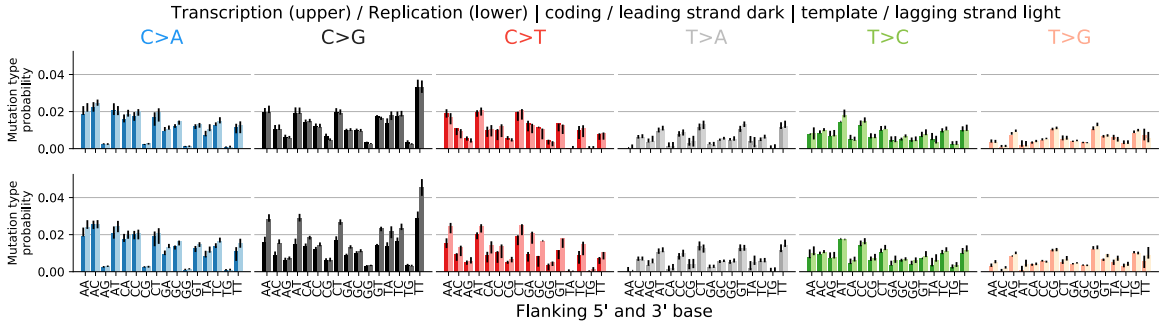


Fig. C.92 **TS19: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

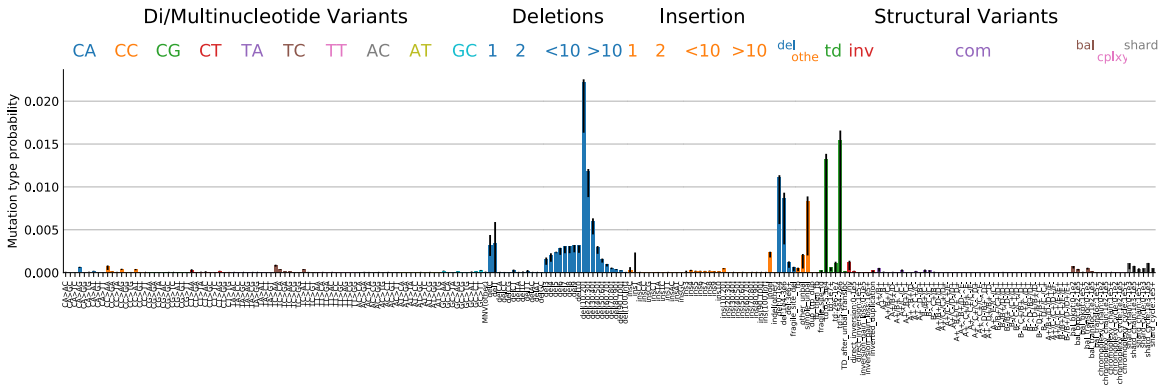


Fig. C.93 **TS19: Spectrum other mutation types.**

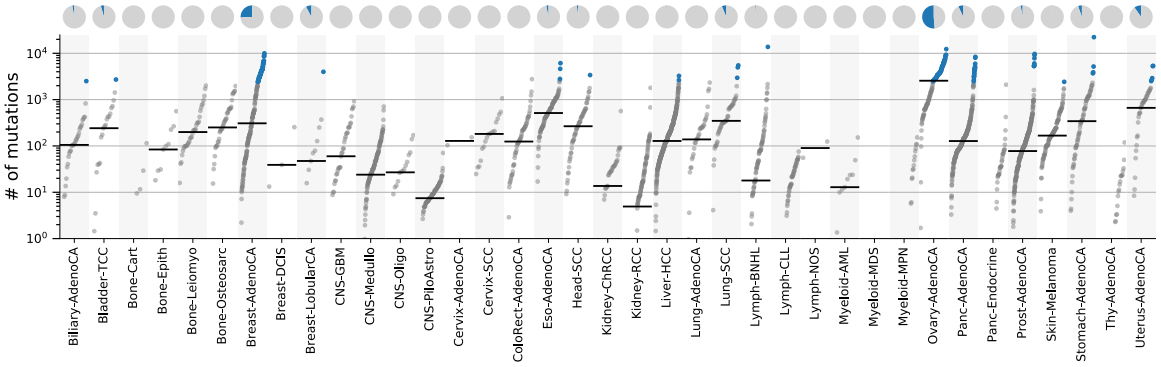


Fig. C.94 **TS19: Signature activity in different cancer types.**

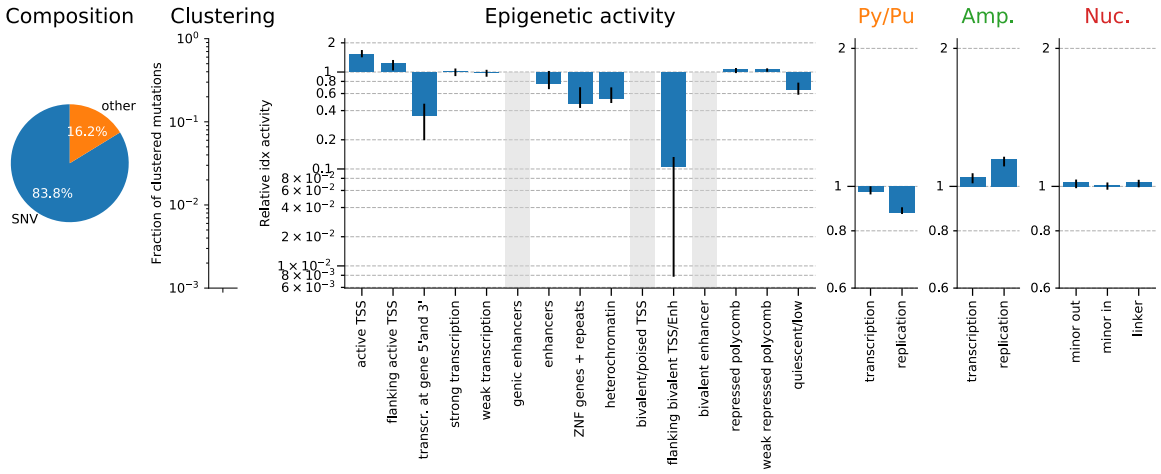


Fig. C.95 **TS19: Signature specific tensor coefficients.**

C.20 TS20-N[T>G]T (unknown/5FU)

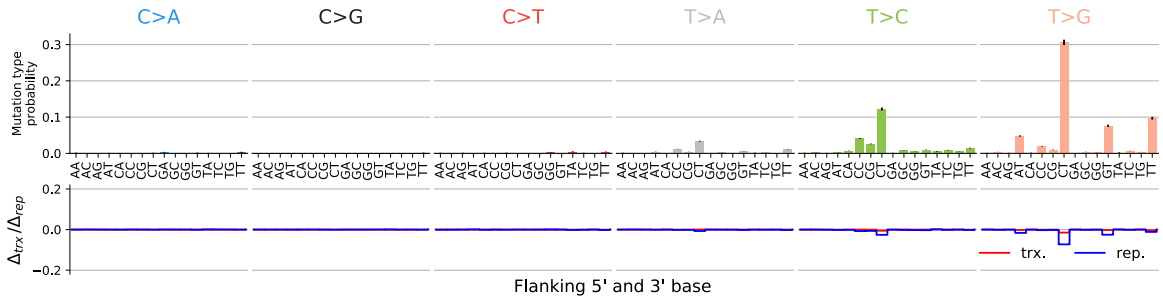


Fig. C.96 **TS20: Single base substitution spectrum.**

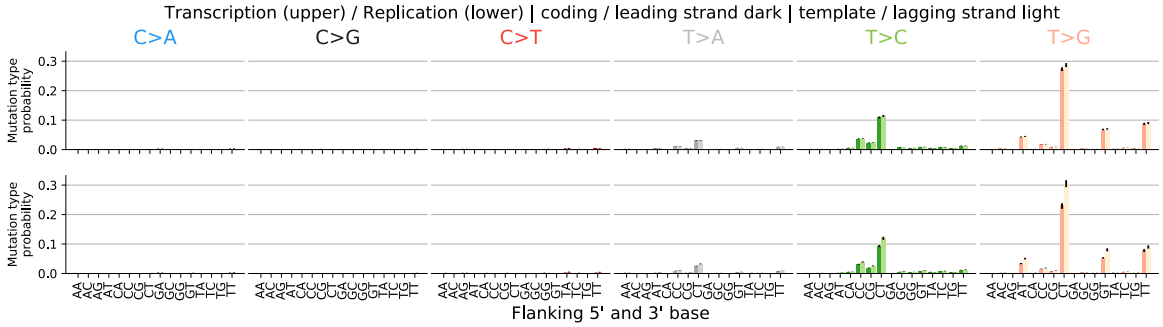


Fig. C.97 **TS20: Single base substitution spectra for template/coding and leading/lagging strand DNA.**

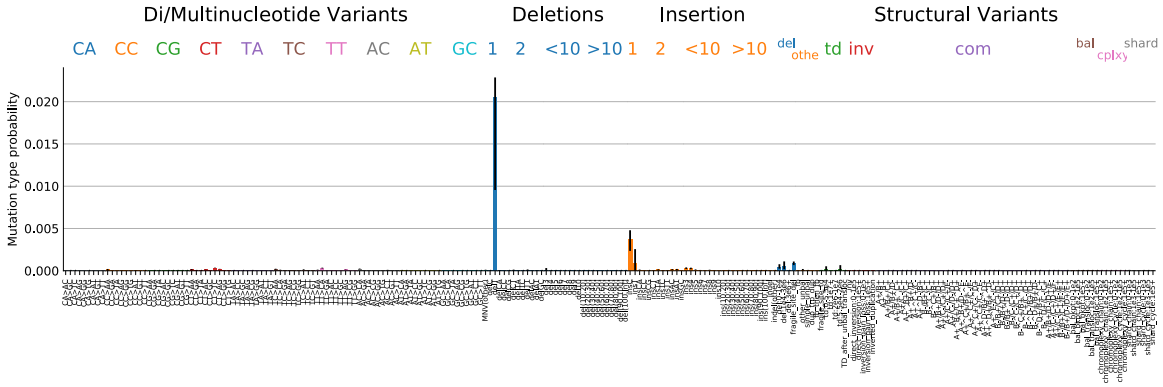


Fig. C.98 **TS20: Spectrum other mutation types.**

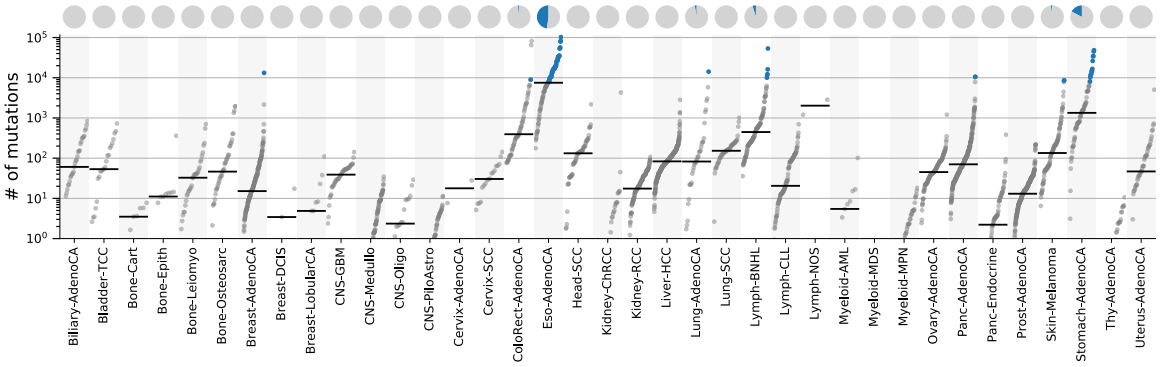


Fig. C.99 **TS20: Signature activity in different cancer types.**

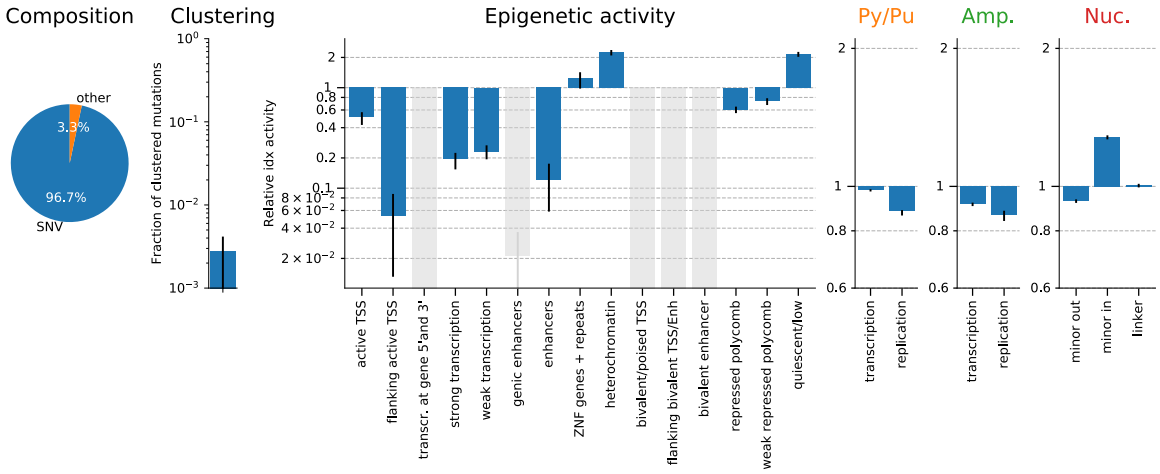


Fig. C.100 **TS20: Signature specific tensor coefficients.**

Appendix D

Additional Analysis

D.1 Elevated somatic mutation burdens in normal human cells due to defective DNA polymerases

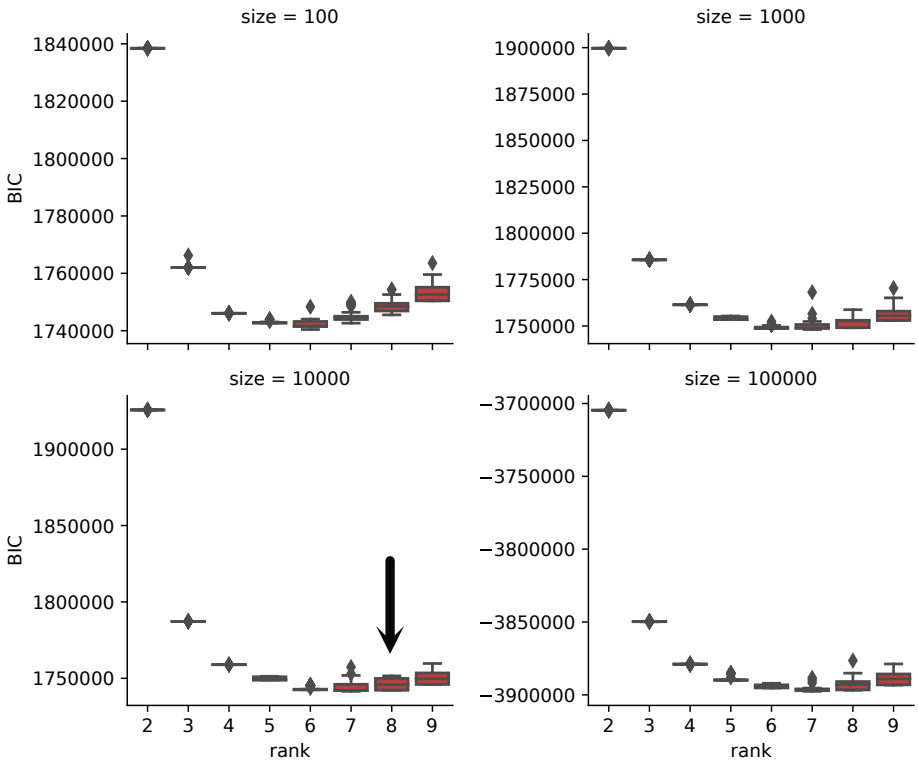


Fig. D.1 **Model selection in the dataset from Robinson et al. (2020).** Chosen number of signatures 8 with a size τ of 10,000.

Fig. D.2 **TS17: Single base substitution spectra for template/coding and leading/lagging strand DNA, as well as the the spectrum for other mutation types.**

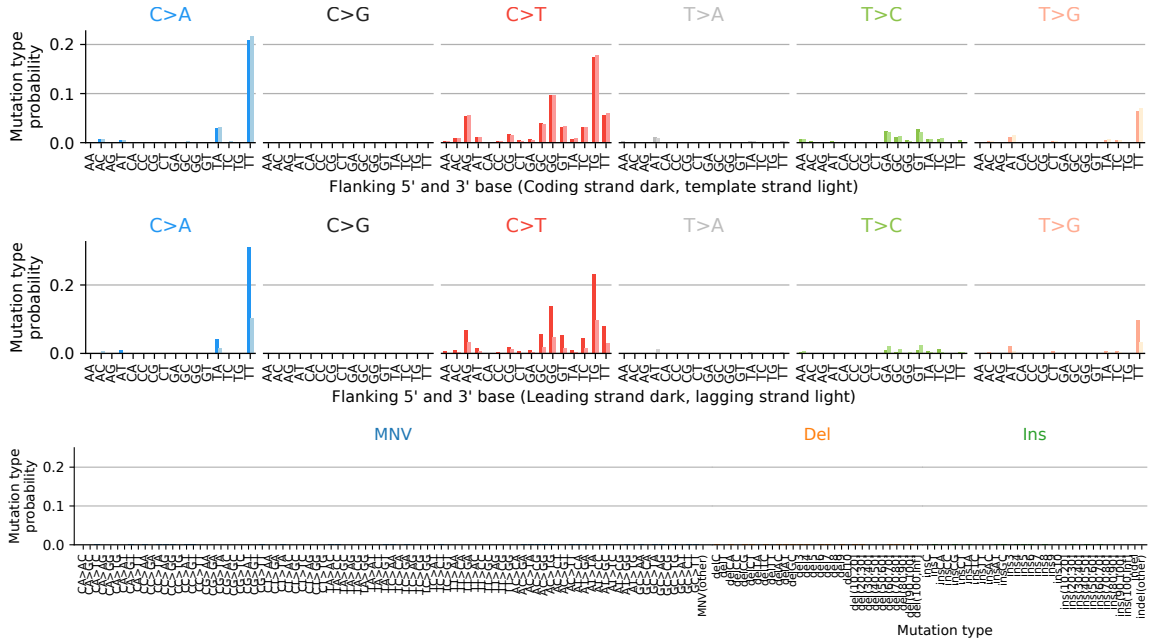


Fig. D.3 **TS17-a: Single base substitution spectra for template/coding and leading/lagging strand DNA, as well as the the spectrum for other mutation types.**

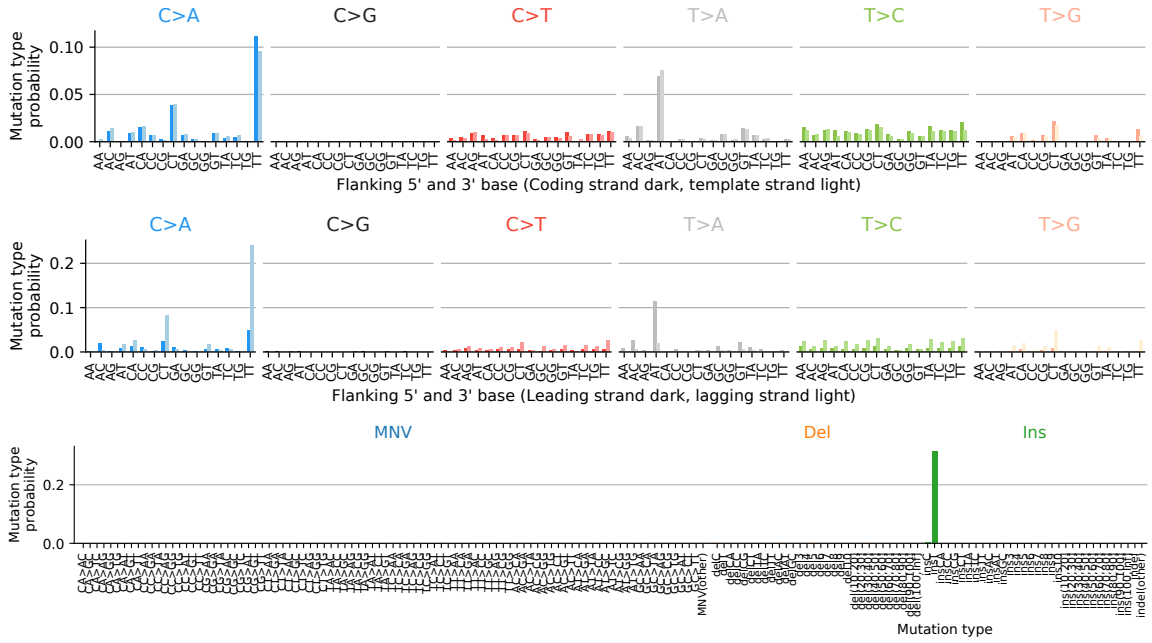


Fig. D.4 **TS-POLD1 (Ins): Single base substitution spectra for template/coding and leading/lagging strand DNA, as well as the the spectrum for other mutation types.**

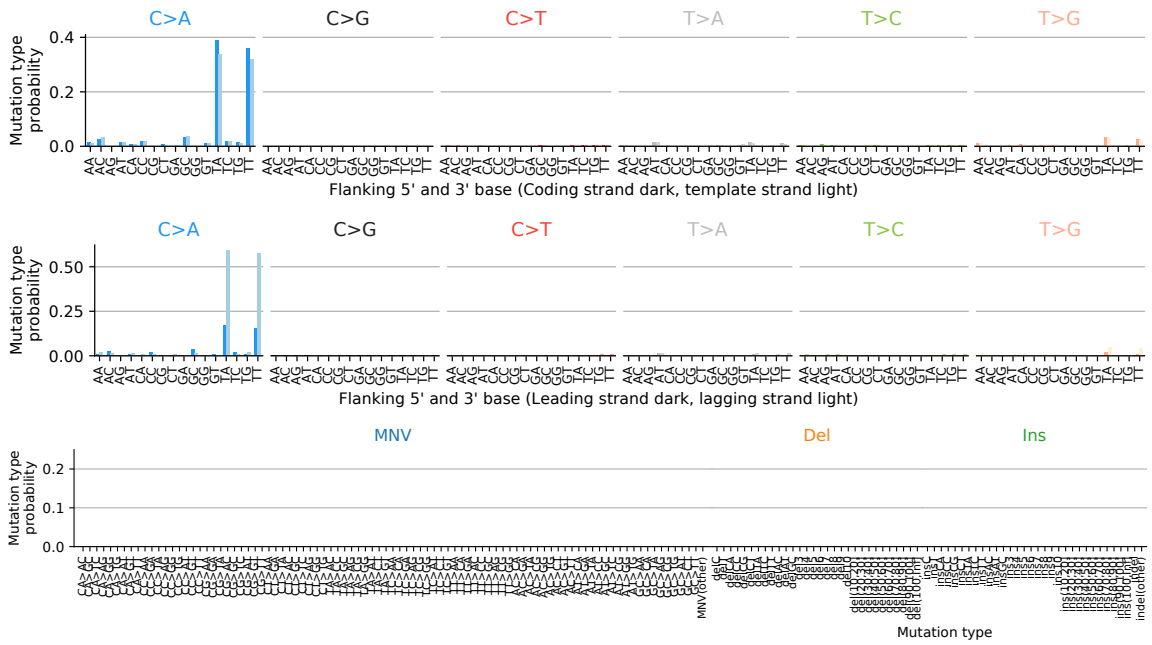


Fig. D.5 **TS-POLD1: Single base substitution spectra for template/coding and leading/lagging strand DNA, as well as the the spectrum for other mutation types.**

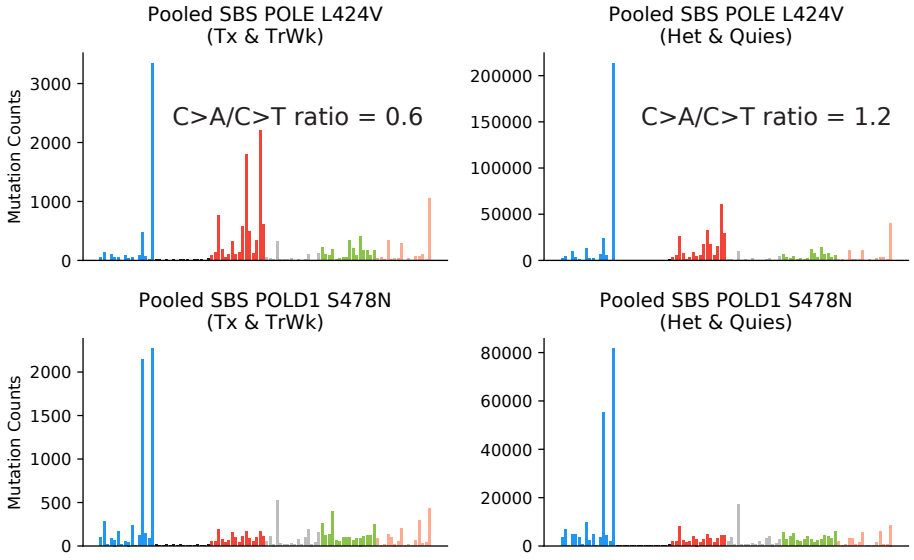


Fig. D.6 **Pooled POLE L424V and POLD1 S478N single base substitutions from active and heterochromatic regions.**

D.2 The mutational signatures of DNA mismatch repair deficiencies

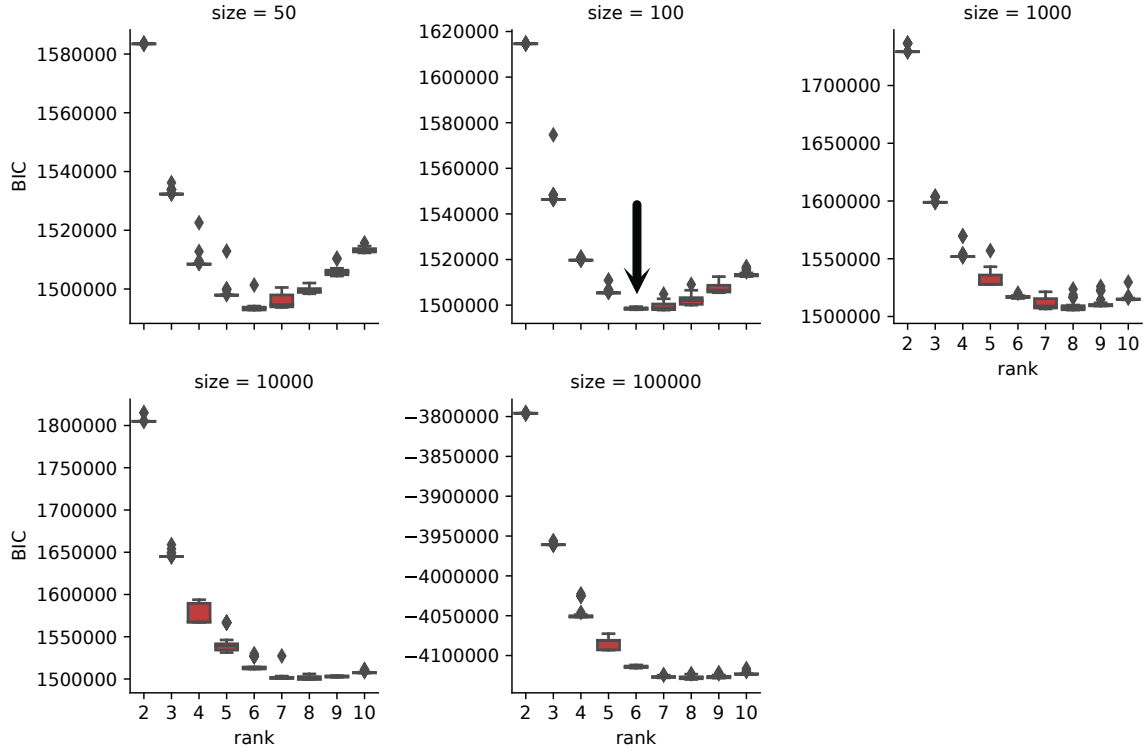


Fig. D.7 **Model selection in the dataset from Mathijs A. Sanders.** Chosen number of signatures 6 with a size τ of 100.

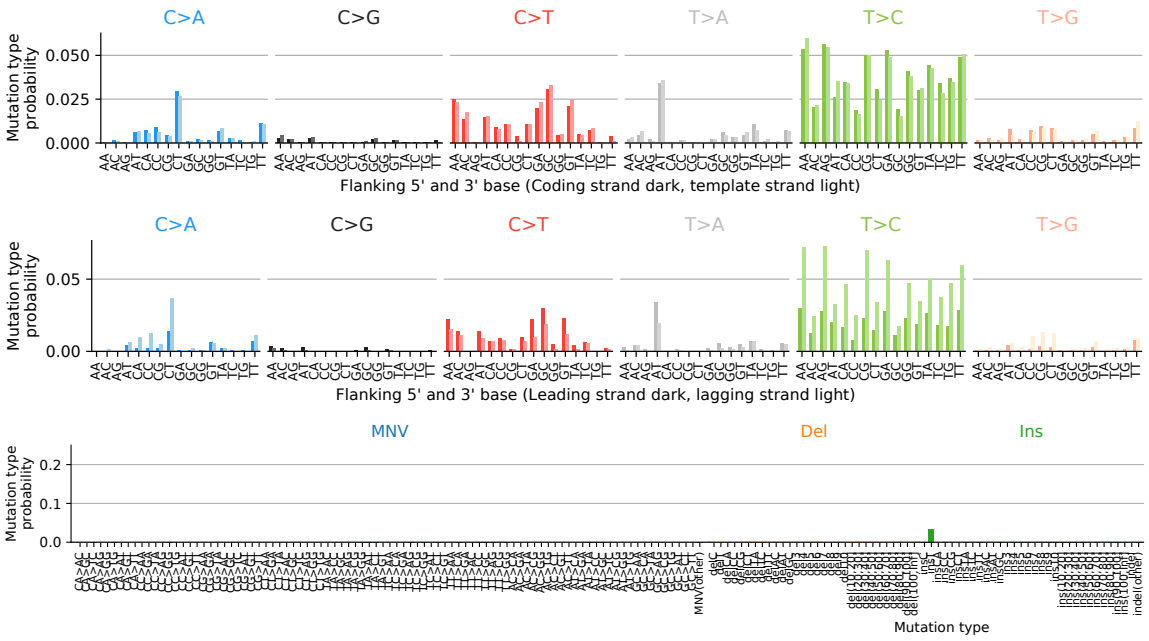


Fig. D.8 **TS-MMRD (T>C): Single base substitution spectra for template/coding and leading/lagging strand DNA, as well as the the spectrum for other mutation types.**

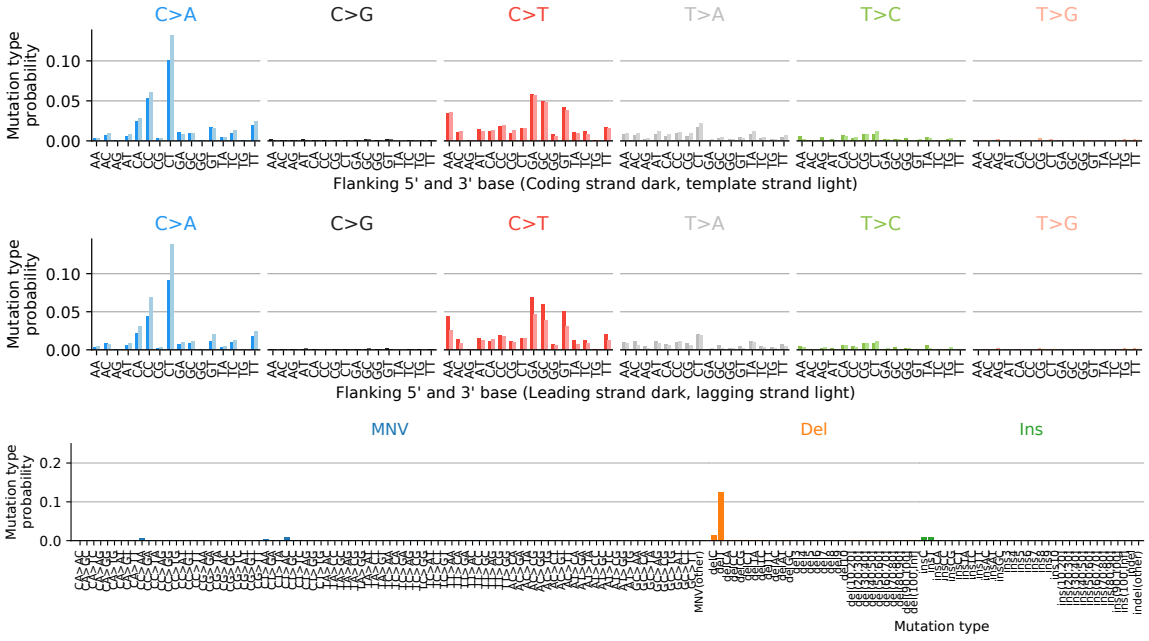


Fig. D.9 **TS-MMRD (C>A): Single base substitution spectra for template/coding and leading/lagging strand DNA, as well as the the spectrum for other mutation types.**

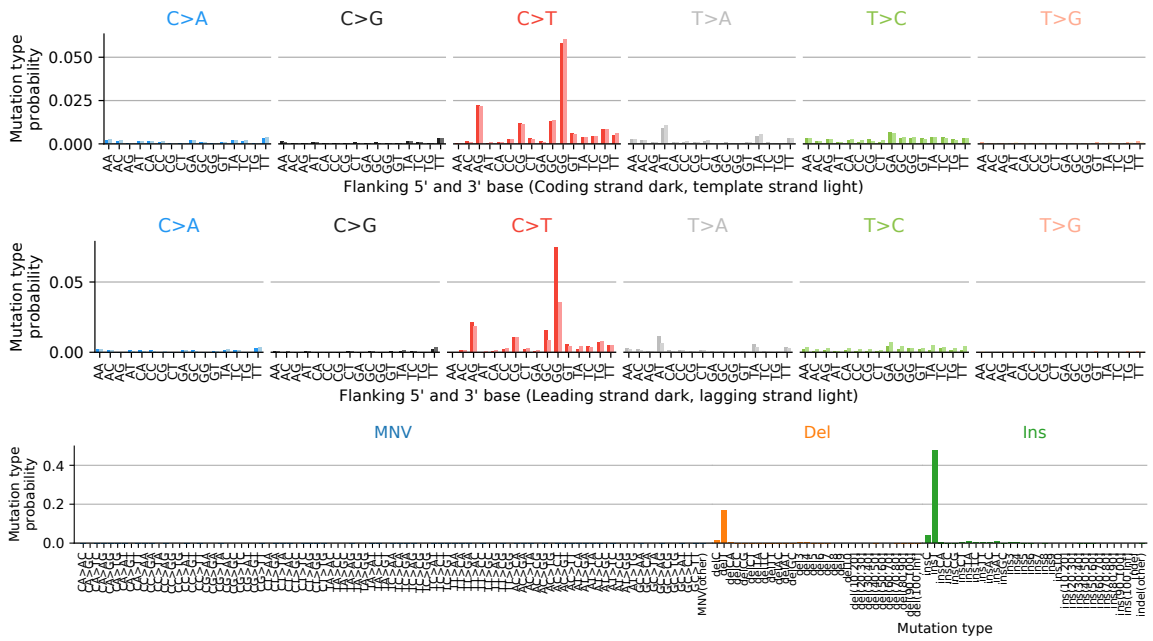


Fig. D.10 **TS-MMRD (Ins): Single base substitution spectra for template/coding and leading/lagging strand DNA, as well as the the spectrum for other mutation types.**

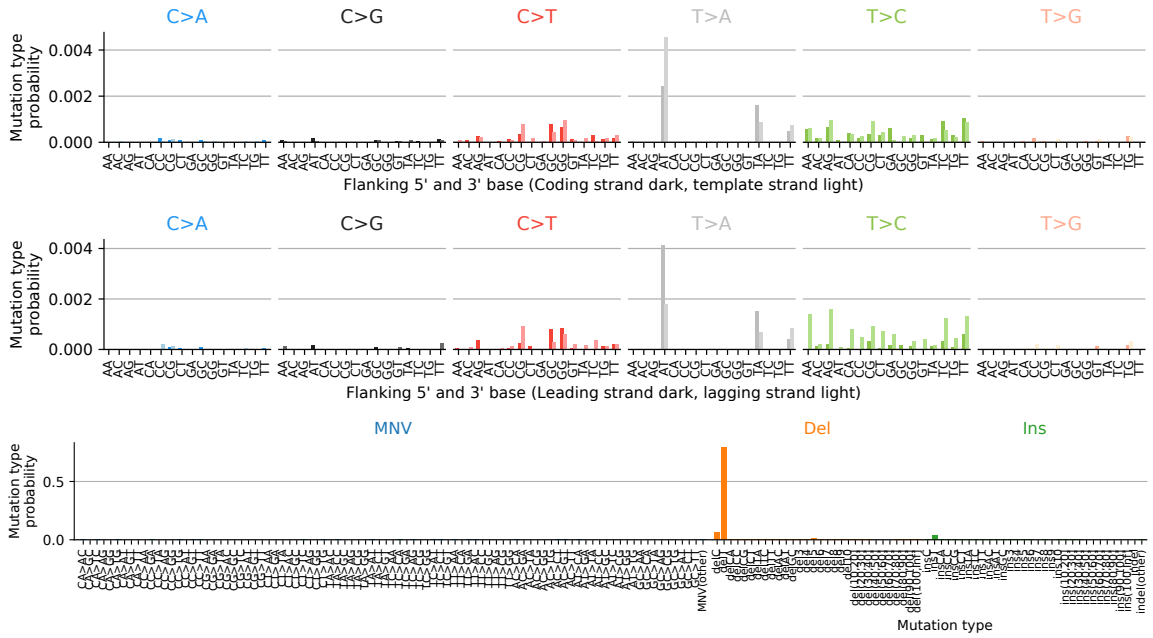


Fig. D.11 **TS-MMRD (Del): Single base substitution spectra for template/coding and leading/lagging strand DNA, as well as the the spectrum for other mutation types.**

D.3 The analysis of the OCCAMS dataset revealed 15 tensor signatures

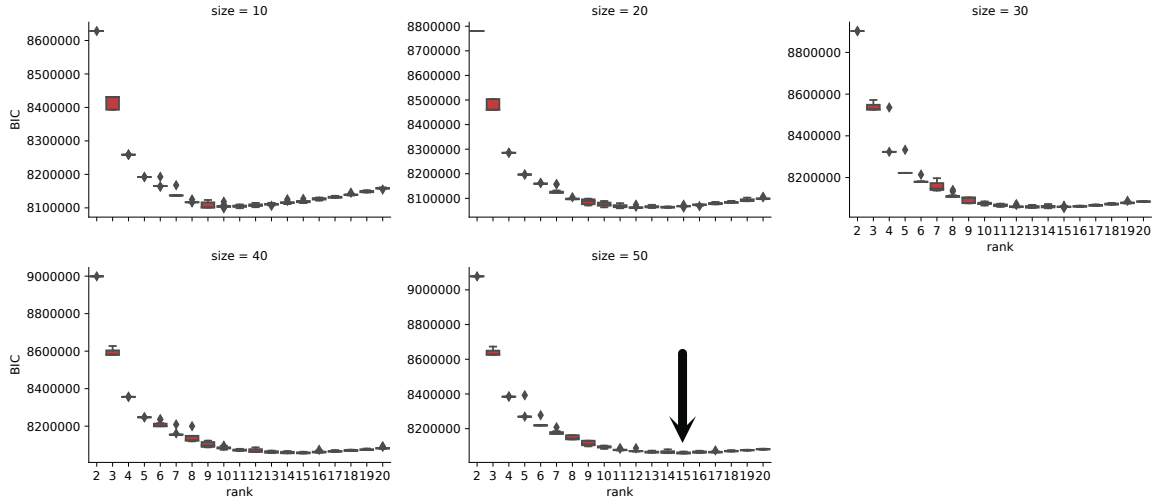


Fig. D.12 **Model selection in the OCCAMS dataset.** Chosen number of signatures 15 with a size τ of 50.

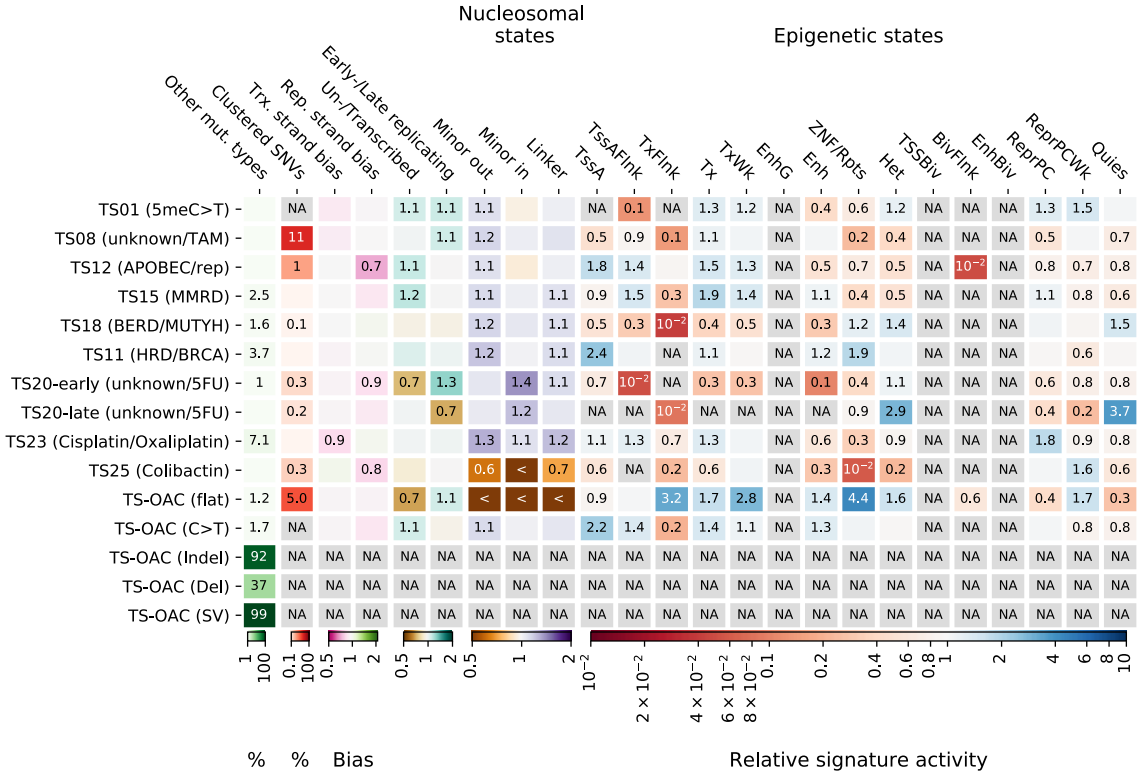


Fig. D.13 **Tensor factors extracted from the OCCAMS dataset (Frankell et al., 2019).** Accompanying tensor factors to the signatures depicted in Fig. 3.29.

D.4 Extensive heterogeneity in somatic mutation and selection in the human bladder

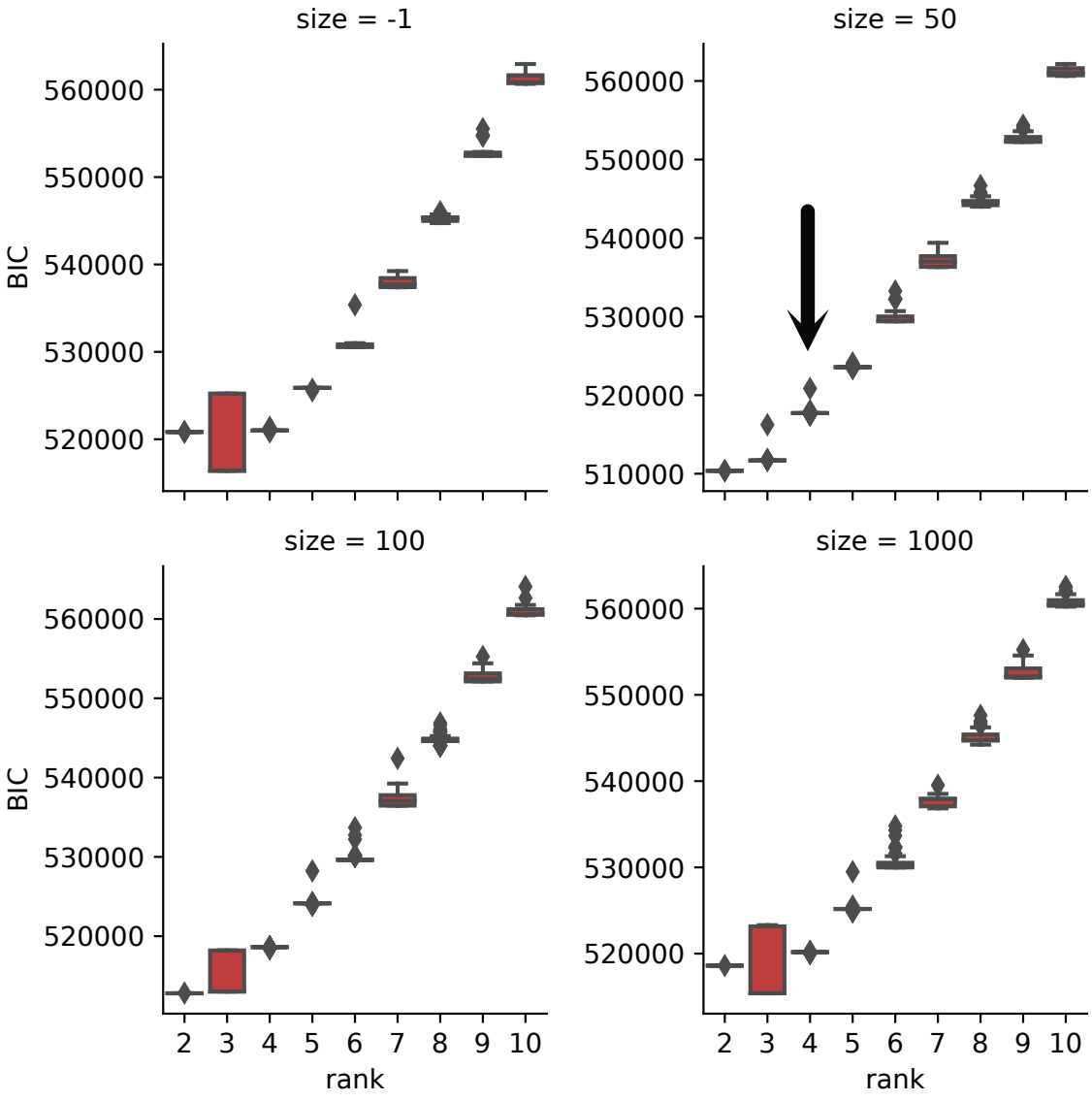


Fig. D.14 **Model selection in the normal bladder urothelium dataset.** Chosen number of signatures 4 with a size τ of 50.

D.5 The mutational processes in normal and tumorous cells

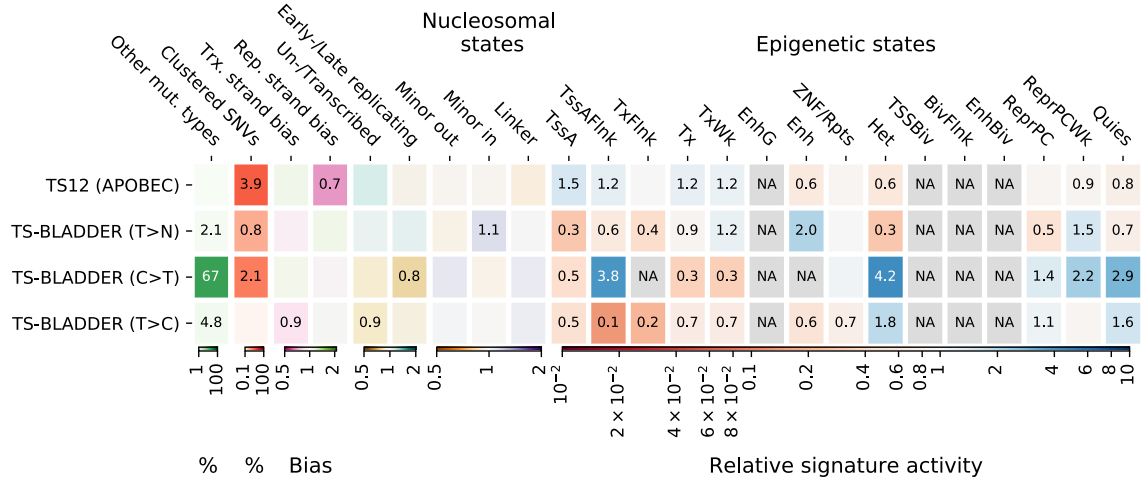


Fig. D.15 **Tensor factors extracted from the normal bladder urothelium dataset.** Accompanying tensor factors to the signatures depicted in Fig. 3.31.

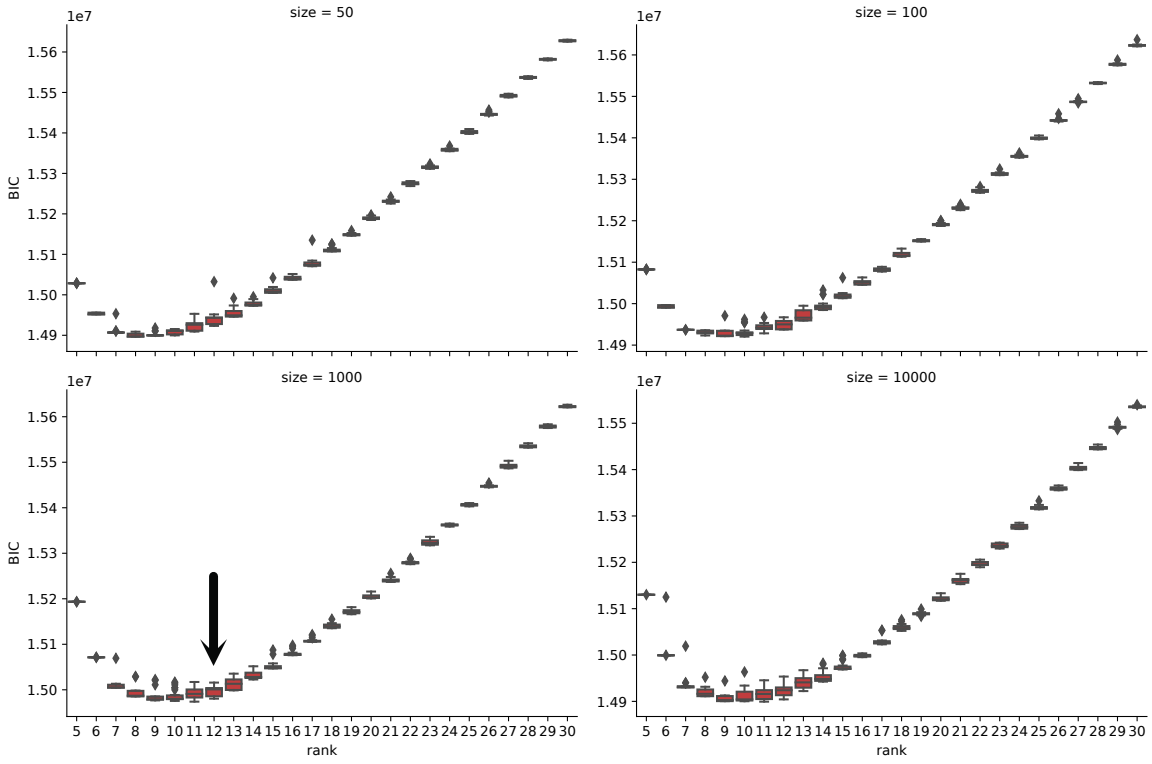


Fig. D.16 **Model selection in the normal and tumour dataset.** Chosen number of signatures 12 with a size τ of 1000.

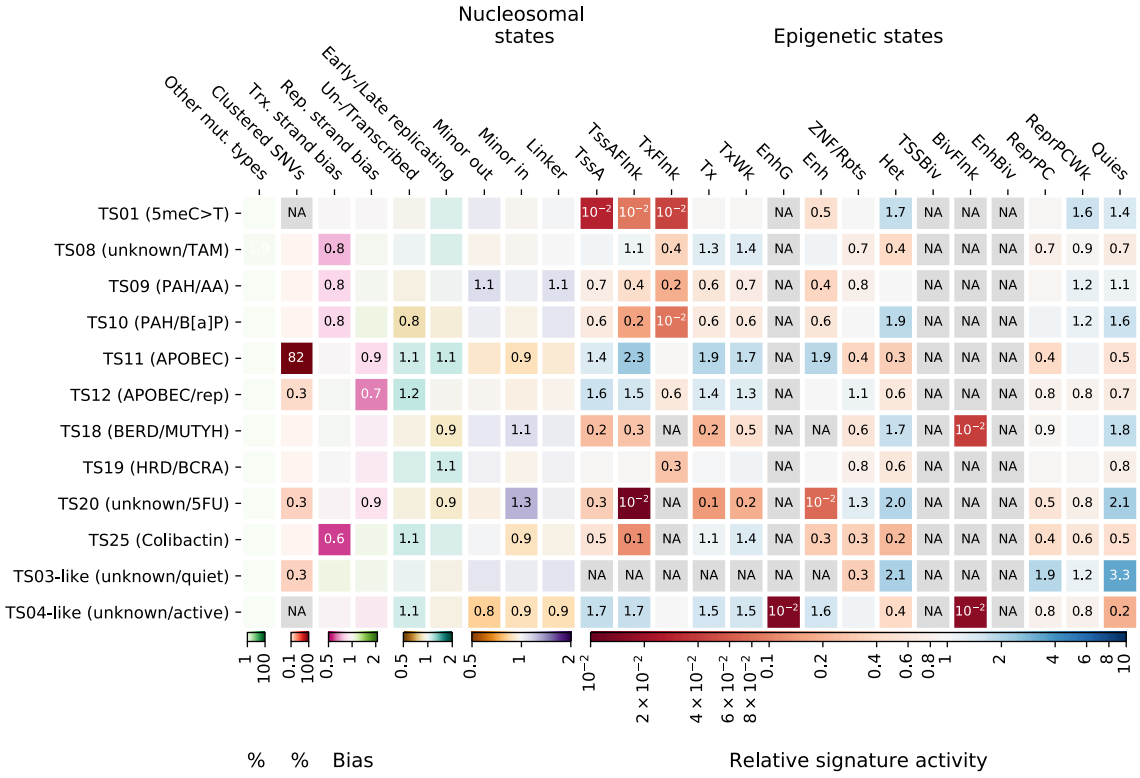


Fig. D.17 **The tensor signatures of normal cells.** The representation of tensor factors analogous to Fig. 3.5.

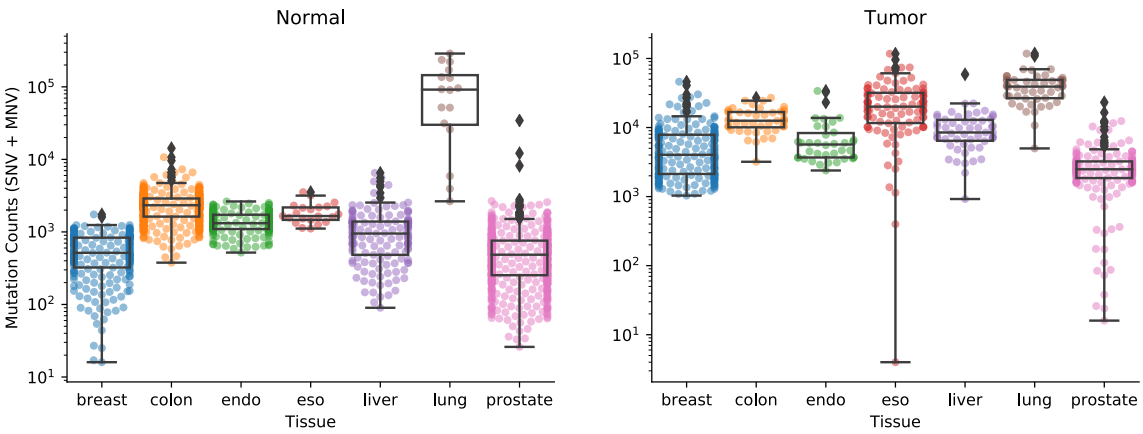


Fig. D.18 **Pooled POLE L424V and POLD1 S478N single base substitutions from active and heterochromatic regions.**

Appendix E

Publications and Conferences

List of publications during PhD studies

1. Yu Fu, Alexander W. Jung, Ramon Viñas Torne, Santiago Gonzalez, **Harald Vöhringer**, Artem Shmatko, Lucy R. Yates, Mercedes Jimenez-Linan, Luiza Moore Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat Cancer 1, 800–810 (2020). https://doi.org/10.1038/s43018-020-0085-8

2. Nadezda V. Volkova, Bettina Meier, Víctor González-Huici, Simone Bertolini, Santiago Gonzalez, **Harald Vöhringer**, Federico Abascal, Iñigo Martincorena, Peter J. Campbell, Anton Gartner Moritz Gerstung. Mutational signatures are jointly shaped by DNA damage and repair. Nat Commun 11, 2169 (2020).

List of submitted manuscripts

1. **Harald Vöhringer**, Arne van Hoeck, Edwin Cuppen, Moritz Gerstung. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. bioRxiv 850453; doi: https://doi.org/10.1101/850453

2. Andrew R. J. Lawson, Federico Abascal, Tim H. H. Coorens, Yvette Hooks, Laura O'Neill, Calli Latimer, Keiran Raine, Mathijs A. Sanders, Anne Y. Warren, Krishnaa T. A. Mahbubani, Bethany Bareham, Timothy M. Butler, Luke M. R. Harvey, Alex Cagan, Andrew Menzies, Luiza Moore, Alexandra J. Colquhoun, William Turner, Benjamin Thomas, Vincent Gnanapragasam, Nicholas Williams, Doris M. Rassl, **Harald Vöhringer**, Sonia Zumalave, Jyoti Nangalia, Jose M. C. Tubio, Moritz Gerstung, Kourosh Saeb-Parsy, Michael R. Stratton, Peter J. Campbell, Thomas J. Mitchell, Inigo

Martincorena. Extensive heterogeneity in somatic mutation and selection in the human bladder.

Attended conferences (posters and talks)

1. The Biology of Genomes, Virtual Conference, May 2020 (Poster)
2. EMBL Cancer Genomics, Heidelberg, November 2019 (Poster)
3. eSCAMPs 2019, Cambridge, February 2019 (Poster)
4. RECOMB 2018, Paris, April 2018 (Talk)
5. Quantitative Genomics, London, 2018 (Poster prize)

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., and Others (2016). Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.

Adelman, K. and Lis, J. T. (2012). Promoter-proximal pausing of rna polymerase ii: emerging roles in metazoans. *Nat Rev Genet*, 13(10):720–731.

Ahnesorg, P., Smith, P., and Jackson, S. P. (2006). Xlf interacts with the xrcc4-dna ligase iv complex to promote dna nonhomologous end-joining. *Cell*, 124(2):301–313.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). *Molecular Biology of the Cell*.

Alexandrov, L., Kim, J., Haradhvala, N. J., Huang, M. N., Ng, A. W. T., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Getz, G., Rozen, S. G., and Stratton, M. R. (2018). The repertoire of mutational signatures in human cancer. *bioRxiv*.

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., and Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nat. Genet.*, 47(12):1402–1407.

Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P. J., Vineis, P., Phillips, D. H., and Stratton, M. R. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622.

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., , Getz, G., Rozen, S. G., and Stratton, M. R. (2019). The repertoire of mutational signatures in human cancer. *bioRxiv*.

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Alexandrov, L. B., Bergstrom, E. N., Boutros, P., Chan, K., Covington, K. R., Fujimoto, A., Getz, G., Gordenin, D. A., Haradhvala, N. J., Islam, S. M. A., Kazanov, M., Klimczak, L. J., Lawrence, M., Martincorena, I., McPherson, J. R., Nakagawa, H.,

Polak, P., Prokopec, S., Roberts, S. A., Rozen, S. G., Saini, N., Shibata, T., Shiraishi, Y., Stratton, M. R., Teh, B. T., Vázquez-García, I., Wheeler, D. A., Yousif, F., Yu, W., Rozen, S. G., Stratton, M. R., Group, P. M. S. W., and Consortium, P. (2020). The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., and et al. (2013a). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013b). Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1):246–259.

Ames, B. N., Mccann, J., and Yamasaki, E. (1975). Methods for detecting carcinogens and mutagens with the salmonella/mammalian-microsome mutagenicity test. *Mutat Res*, 31(6):347–364.

Amouroux, R., Campalans, A., Epe, B., and Radicella, J. P. (2010). Oxidative stress triggers the preferential assembly of base excision repair complexes on open chromatin regions. *Nucleic Acids Research*, 38(9):2878–2890

Ayoub, N., Jeyasekharan, A. D., Bernal, J. A., and Venkitaraman, A. R. (2008). Hp1-βmobilization promotes chromatin changes that initiate the dna damage response. *Nature*, 453(7195):682–686.

Bae, S. H., Bae, K. H., Kim, J. A., and Seo, Y. S. (2001). Rpa governs endonuclease switching during processing of okazaki fragments in eukaryotes. *Nature*, 412(6845):456–461.

Baird, W. M., Hooven, L. A., and Mahadevan, B. (2005). Carcinogenic polycyclic aromatic hydrocarbon-dna adducts and mechanism of action. *Environ Mol Mutagen*, 45(2-3):106–114.

Bell, O., Tiwari, V. K., Thomä, N. H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. *Nature Reviews Genetics*, 12(8):554–564.

Bergstrom, E. N., Barnes, M., Martincorena, I., and Alexandrov, L. B. (2020). Generating realistic null hypothesis of cancer mutational landscapes using sigprofilersimulator. *bioRxiv*.

Bergstrom, E. N., Huang, M. N., Mahto, U., Barnes, M., Stratton, M. R., Rozen, S. G., and Alexandrov, L. B. (2019). Sigprofilermatrixgenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics*, 20(1):685.

Bhattacharya, S. K., Ramchandani, S., Cervoni, N., and Szyf, M. (1999). A mammalian protein with specific demethylase activity for mcpg dna. *Nature*, 397(6720):579–583.

Boffetta, P., Jourenkova, N., and Gustavsson, P. (1997). Cancer risk from occupational and environmental exposure to polycyclic aromatic hydrocarbons. *Cancer Causes & Control*, 8(3):444–472.

Borrego-Soto, G., Ortiz-López, R., and Rojas-Martínez, A. (2015). Ionizing radiation-induced dna injury and damage detection in patients with breast cancer. *Genet Mol Biol*, 38(4):420–432.

Boström, C.-E., Gerde, P., Hanberg, A., Jernström, B., Johansson, C., Kyrklund, T., Rannug, A., Törnqvist, M., Victorin, K., and Westerholm, R. (2002). Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. *Environmental health perspectives*, 110 Suppl 3(Suppl 3):451–488.

Braithwaite, E. K., Prasad, R., Shock, D. D., Hou, E. W., Beard, W. A., and Wilson, S. H. (2005). Dna polymerase lambda mediates a back-up base excision repair activity in extracts of mouse embryonic fibroblasts. *J Biol Chem*, 280(18):18469–18475.

Britton, S., Coates, J., and Jackson, S. P. (2013). A new method for high-resolution imaging of ku foci to decipher mechanisms of dna double-strand break repair. *The Journal of cell biology*, 202(3):579–595.

Brunner, S. F., Roberts, N. D., Wylie, L. A., Moore, L., Aitken, S. J., Davies, S. E., Sanders, M. A., Ellis, P., Alder, C., Hooks, Y., Abascal, F., Stratton, M. R., Martincorena, I., Hoare, M., and Campbell, P. J. (2019). Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature*, 574(7779):538–542.

Bzymek, M., Thayer, N. H., Oh, S. D., Kleckner, N., and Hunter, N. (2010). Double holliday junctions are intermediates of dna break repair. *Nature*, 464(7290):937–941.

Cerritelli, S. M. and Crouch, R. J. (2009). Ribonuclease h: the enzymes in eukaryotes. *FEBS J*, 276(6):1494–1505.

Cerritelli, S. M. and Crouch, R. J. (2016). The balancing act of ribonucleotides in dna. *Trends in biochemical sciences*, 41(5):434–445.

Champoux, J. J. (2001). Dna topoisomerases: Structure, function, and mechanism. *Annual Review of Biochemistry*, 70(1):369–413.

Chan, K., Roberts, S. A., Klimczak, L. J., Sterling, J. F., Saini, N., Malc, E. P., Kim, J., Kwiatkowski, D. J., Fargo, D. C., Mieczkowski, P. A., Getz, G., and Gordenin, D. A. (2015). An apobec3a hypermutation signature is distinguishable from the signature of background mutagenesis by apobec3b in human cancers. *Nat Genet*, 47(9):1067–1072.

Chapman, M. A., Lawrence, M. S., Keats, J. J., Cibulskis, K., Sougnez, C., Schinzel, A. C., Harview, C. L., Brunet, J.-P., Ahmann, G. J., Adli, M., Anderson, K. C., Ardlie, K. G., Auclair, D., Baker, A., Bergsagel, P. L., Bernstein, B. E., Drier, Y., Fonseca, R., Gabriel, S. B., Hofmeister, C. C., Jagannath, S., Jakubowiak, A. J., Krishnan, A., Levy, J., Liefeld, T., Lonial, S., Mahan, S., Mfuko, B., Monti, S., Perkins, L. M., Onofrio, R., Pugh, T. J., Rajkumar, S. V., Ramos, A. H., Siegel, D. S., Sivachenko, A., Stewart, A. K., Trudel, S., Vij, R., Voet, D., Winckler, W., Zimmerman, T., Carpten, J., Trent, J., Hahn, W. C., Garraway, L. A., Meyerson, M., Lander, E. S., Getz, G., and Golub, T. R. (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339):467–472.

Chen, C.-L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., d'Aubenton Carafa, Y., Arneodo, A., Hyrien, O., and Thermes, C. (2010). Impact of replication timing on non-cpg and cpg substitution rates in mammalian genomes. *Genome Res*, 20(4):447–457.

Chen, P., Zhao, J., Wang, Y., Wang, M., Long, H., Liang, D., Huang, L., Wen, Z., Li, W., Li, X., Feng, H., Zhao, H., Zhu, P., Li, M., Wang, Q.-f., and Li, G. (2013). H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes & development*, 27(19):2109–2124.

Chen, X., Chen, Z., Chen, H., Su, Z., Yang, J., Lin, F., Shi, S., and He, X. (2012). Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science*, 335(6073):1235–1238.

Christensen, S., vd Roest, B., Besselink, N., Janssen, R., Boymans, S., Martens, J., Yaspo, M.-L., Priestley, P., Kuijk, E., Cuppen, E., et al. (2019). 5-fluorouracil treatment induces characteristic t> g mutations in human cancer. *bioRxiv*, page 681262.

Chueh, A. C., Wong, L. H., Wong, N., and Choo, K. H. A. (2005). Variable and hierarchical size distribution of l1-retroelement-enriched cenp-a clusters within a functional human neocentromere. *Hum Mol Genet*, 14(1):85–93.

De, S. and Michor, F. (2011). Dna secondary structures and epigenetic determinants of cancer genome evolution. *Nature structural & molecular biology*, 18(8):950–955.

De Bont, R. and Van Larebeke, N. (2004). Endogenous dna damage in humans: a review of quantitative data. *Mutagenesis*, 19(3):169–185.

Dennehey, B. K. and Tyler, J. (2014). *Histone Chaperones in the Assembly and Disassembly of Chromatin*, pages 29–67. Springer New York, New York, NY.

Devasagayam, T., Tilak, J., Boloor, K., Sane, K. S., Ghaskadbi, S. S., and Lele, R. (2004). Free radicals and antioxidants in human health: current status and future prospects. *Japi*, 52(794804):4.

Donahue, B. A., Yin, S., Taylor, J. S., Reines, D., and Hanawalt, P. C. (1994). Transcript cleavage by rna polymerase ii arrested by a cyclobutane pyrimidine dimer in the dna template. *Proceedings of the National Academy of Sciences*, 91(18):8502.

Doublié, S. and Zahn, K. E. (2014). Structural insights into eukaryotic dna replication. *Front Microbiol*, 5:444.

Drost, J., van Boxtel, R., Blokzijl, F., Mizutani, T., Sasaki, N., Sasselli, V., de Ligt, J., Behjati, S., Grolleman, J. E., van Wezel, T., Nik-Zainal, S., Kuiper, R. P., Cuppen, E., and Clevers, H. (2017). Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science*, 358(6360):234–238.

Duncan, B. K. and Weiss, B. (1982). Specific mutator effects of ung (uracil-dna glycosylase) mutations in escherichia coli. *Journal of bacteriology*, 151(2):750–755.

Eberharter, A. and Becker, P. B. (2002). Histone acetylation: a switch between repressive and permissive chromatin. second in review series on chromatin dynamics. *EMBO reports*, 3(3):224–229.

Enemark, E. J. and Joshua-Tor, L. (2006). Mechanism of DNA translocation in a replicative hexameric helicase. *Nature*, 442(7100):270–275.

Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nature Methods*, 9:215 EP –.

Fagbemi, A. F., Orelli, B., and Schärer, O. D. (2011). Regulation of endonuclease activity in human nucleotide excision repair. *DNA repair*, 10(7):722–729.

Ferguson, L. R. and Denny, W. A. (2007). Genotoxicity of non-covalent interactions: Dna intercalators. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 623(1):14–23.

Fischer, A., Illingworth, C. J., Campbell, P. J., and Mustonen, V. (2013). Emu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome biology*, 14(4):R39.

Fishel, R. and Lee, J.-B. (2016). *Mismatch Repair*, pages 305–339. Springer Japan, Tokyo.

Flores-Rozas, H., Clark, D., and Kolodner, R. D. (2000). Proliferating cell nuclear antigen and msh2p-msh6p interact to form an active mispair recognition complex. *Nat Genet*, 26(3):375–378.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U., and Campbell, P. J. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(D1):D805–D811.

Frankell, A. M., Jammula, S., Li, X., Contino, G., Killcoyne, S., Abbas, S., Perner, J., Bower, L., Devonshire, G., Ococks, E., Grehan, N., Mok, J., O'Donovan, M., MacRae, S., Eldridge, M. D., Tavaré, S., Fitzgerald, R. C., Noorani, A., Edwards, P. A. W., Nutzinger, B., Hughes, C., Fidziukiewicz, E., Northrop, A., de la Rue, R., Katz-Summercorn, A., Loureda, D., Miremadi, A., Malhotra, S., Tripathi, M., Lynch, A. G., Eldridge, M., Secrier, M., Davies, J., Crichton, C., Carroll, N., Safranek, P., Hindmarsh, A., Sujendran, V., Hayes, S. J., Ang, Y., Sharrocks, A., Preston, S. R., Oakes, S., Bagwan, I., Save, V., Skipworth, R. J. E., Hupp, T. R., O'Neill, J. R., Tucker, O., Beggs, A., Taniere, P., Puig, S., Underwood, T. J., Walker, R. C., Grace, B. L., Barr, H., Shepherd, N., Old, O., Lagergren, J., Gossage, J., Davies, A., Chang, F., Zylstra, J., Mahadeva, U., Goh, V., Ciccarelli, F. D., Sanders, G., Berrisford, R., Harden, C., Lewis, M., Cheong, E., Kumar, B., Parsons, S. L., Soomro, I., Kaye, P., Saunders, J., Lovat, L., Haidry, R., Igali, L., Scott, M., Sothi, S., Suortamo, S., Lishman, S., Hanna, G. B., Moorthy, K., Peters, C. J., Grabowska, A., Turkington, R., McManus, D., Coleman, H., Khoo, D., Fickling, W., Fitzgerald, R. C., the Oesophageal Cancer Clinical, and Consortium, M. S. O. (2019). The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. *Nature Genetics*, 51(3):506–516.

Frederico, L. A., Kunkel, T. A., and Shaw, B. R. (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*, 29(10):2532–2537.

Frick, D. and Richardson, C. (2001). Dna primases. *Annual review of biochemistry*, 70:39–80.

Frigola, J., Sabarinathan, R., Mularoni, L., Muiños, F., Gonzalez-Perez, A., and López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. *Nature Genetics*, 49(12):1684–1692.

Gaszner, M. and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet*, 7(9):703–713.

Goodrich, J. A. and Tjian, R. (1994). Transcription factors iie and iih and atp hydrolysis direct promoter clearance by rna polymerase ii. *Cell*, 77(1):145–156.

Gottlieb, T. M. and Jackson, S. P. (1993). The dna-dependent protein kinase: requirement for dna ends and association with ku antigen. *Cell*, 72(1):131–142.

Gregory, T. R. (2011). *The evolution of the genome*. Elsevier.

Guhaniyogi, J. and Brewer, G. (2001). Regulation of mrna stability in mammalian cells. *Gene*, 265(1-2):11–23.

Ha, T. (2007). Need for speed: Mechanical regulation of a replicative helicase. *Cell*, 129(7):1249–1250.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674.

Hanawalt, P. C. and Spivak, G. (2008). Transcription-coupled dna repair: two decades of progress and surprises. *Nature reviews Molecular cell biology*, 9(12):958.

Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., Dorschner, M. O., Gartler, S. M., and Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated dna reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1):139–144.

Haradhvala, N. J., Kim, J., Maruvka, Y. E., Polak, P., Rosebrock, D., Livitz, D., Hess, J. M., Leshchiner, I., Kamburov, A., Mouw, K. W., Lawrence, M. S., and Getz, G. (2018). Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.*, 9(1):1746.

Haradhvala, N. J., Polak, P., Stojanov, P., Covington, K. R., Shinbrot, E., Hess, J. M., Rheinbay, E., Kim, J., Maruvka, Y. E., Braunstein, L. Z., Kamburov, A., Hanawalt, P. C., Wheeler, D. A., Koren, A., Lawrence, M. S., and Getz, G. (2016). Mutational strand asymmetries in cancer genomes reveal mechanisms of dna damage and repair. *Cell*, 164(3):538 – 549.

Hatch, C. L. and Bonner, W. M. (1990). The human histone h2a.z gene. sequence and regulation. *J Biol Chem*, 265(25):15211–15218.

Hayward, N. K., Wilmott, J. S., Waddell, N., Johansson, P. A., Field, M. A., Nones, K., Patch, A.-M., Kakavand, H., Alexandrov, L. B., Burke, H., and et al. (2017). Whole-genome landscapes of major melanoma subtypes. *Nature*, 545(7653):175–180.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, 15(9):585–598.

Heller, R. C. and Marians, K. J. (2006). Replisome assembly and the direct restart of stalled replication forks. *Nat Rev Mol Cell Biol*, 7(12):932–943.

Hendel, A., Ziv, O., Gueranger, Q., Geacintov, N., and Livneh, Z. (2008). Reduced efficiency and increased mutagenicity of translesion dna synthesis across a tt cyclobutane pyrimidine dimer, but not a tt 6-4 photoproduct, in human cells lacking dna polymerase η. *DNA Repair*, 7(10):1636–1646.

Heyer, W.-D., Ehmsen, K. T., and Liu, J. (2010). Regulation of homologous recombination in eukaryotes. *Annual Review of Genetics*, 44(1):113–139.

Higuchi, Y. and Linn, S. (1995). Purification of all forms of hela cell mitochondrial dna and assessment of damage to it caused by hydrogen peroxide treatment of mitochondria or cells. *J Biol Chem*, 270(14):7950–7956.

Hsieh, P. and Yamane, K. (2008). Dna mismatch repair: molecular mechanism, cancer, and ageing. *Mechanisms of ageing and development*, 129(7-8):391–407.

Hsin, J.-P. and Manley, J. L. (2012). The rna polymerase ii ctd coordinates transcription and rna processing. *Genes & development*, 26(19):2119–2137.

Huang, Y., Fang, J., Bedford, M. T., Zhang, Y., and Xu, R.-M. (2006). Recognition of histone h3 lysine-4 methylation by the double tudor domain of jmjd2a. *Science*, 312(5774):748–751.

Hublitz, P., Albert, M., Hfmpeters, A., Hublitz, P., Albert, M., and Peters, A. H. (2009). Mechanisms of transcriptional repression by histone lysine methylation. *The International Journal of Developmental Biology*, 53(2-3):335–354.

Hübscher, U., Maga, G., and Spadari, S. (2002). Eukaryotic dna polymerases. *Annual Review of Biochemistry*, 71(1):133–163.

Hyun, K., Jeon, J., Park, K., and Kim, J. (2017). Writing, erasing and reading histone lysine methylations. *Experimental & Molecular Medicine*, 49(4):e324–e324.

Ikehata, H. and Ono, T. (2011). The mechanisms of uv mutagenesis. *Journal of radiation research*, 52(2):115–125.

Imielinski, M., Guo, G., and Meyerson, M. (2017). Insertions and deletions target lineage-defining genes in human cancers. *Cell*, 168(3):460–472.

Irimia, A., Eoff, R. L., Guengerich, F. P., and Egli, M. (2009). Structural and functional elucidation of the mechanism promoting error-prone synthesis by human dna polymerase kappa opposite the 7,8-dihydro-8-oxo-2'-deoxyguanosine adduct. *The Journal of biological chemistry*, 284(33):22467–22480.

Iwase, S., Lan, F., Bayliss, P., de la Torre-Ubieta, L., Huarte, M., Qi, H. H., Whetstine, J. R., Bonni, A., Roberts, T. M., and Shi, Y. (2007). The x-linked mental retardation gene smcx/jarid1c defines a family of histone h3 lysine 4 demethylases. *Cell*, 128(6):1077–1088.

Iyer, R. R., Pluciennik, A., Burdett, V., and Modrich, P. L. (2006). Dna mismatch repair: Functions and mechanisms. *Chemical Reviews*, 106(2):302–323.

Iyer, R. R., Pohlhaus, T. J., Chen, S., Hura, G. L., Dzantiev, L., Beese, L. S., and Modrich, P. (2008). The mutsalpha-proliferating cell nuclear antigen interaction in human dna mismatch repair. *J Biol Chem*, 283(19):13310–13319.

Jeffreys, A. J., Royle, N. J., Wilson, V., and Wong, Z. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human dna. *Nature*, 332(6161):278–281.

Jenuwein, T. (2001). Translating the histone code. *Science*, 293(5532):1074–1080.

Jinks-Robertson, S. and Bhagwat, A. S. (2014). Transcription-associated mutagenesis. *Annual Review of Genetics*, 48(1):341–359.

Johnson, K. A. (1993). Conformational coupling in dna polymerase fidelity. *Annual Review of Biochemistry*, 62(1):685–713.

Joyce, C. M. and Steitz, T. A. (1994). Function and structure relationships in dna polymerases. *Annual Review of Biochemistry*, 63(1):777–822.

Kadyrov, F. A., Dzantiev, L., Constantin, N., and Modrich, P. (2006). Endonucleolytic function of MutLalpha in human mismatch repair. *Cell*, 126(2):297–308.

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J., and Segal, E. (2009). The dna-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366.

Karran, P. and Attard, N. (2008). Thiopurines in current medical practice: molecular mechanisms and contributions to therapy-related cancer. *Nature Reviews Cancer*, 8(1):24–36.

Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., Haradhvala, N., Lawrence, M. S., Kiezun, A., Fernandes, S. M., Bahl, S., and et al. (2015). Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature Communications*, 6(1).

Kaufman, E. R. (1984). Replication of dna containing 5-bromouracil can be mutagenic in syrian hamster cells. *Molecular and cellular biology*, 4(11):2449–2454.

Kew, M. C. (2013). Aflatoxins as a cause of hepatocellular carcinoma. *Journal of Gastrointestinal & Liver Diseases*, 22(3).

Kim, J., Mouw, K. W., Polak, P., Braunstein, L. Z., Kamburov, A., Kwiatkowski, D. J., Rosenberg, J. E., Van Allen, E. M., D'Andrea, A., and Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.*, 48(6):600–606.

Kim, M., Krogan, N. J., Vasiljeva, L., Rando, O. J., Nedea, E., Greenblatt, J. F., and Buratowski, S. (2004). The yeast rat1 exonuclease promotes transcription termination by rna polymerase ii. *Nature*, 432(7016):517–522.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., and Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187.

Klemm, S. L., Shipony, Z., and Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220.

Knobel, P. A. and Marti, T. M. (2011). Translesion DNA synthesis in the context of cancer research. *Cancer Cell Int.*, 11:39.

Kondo, N., Takahashi, A., Ono, K., and Ohnishi, T. (2010). Dna damage induced by alkylating agents and repair pathways. *Journal of nucleic acids*, 2010.

Koren, A., Polak, P., Nemesh, J., Michaelson, J. J., Sebat, J., Sunyaev, S. R., and McCarroll, S. A. (2012). Differential relationship of dna replication timing to different forms of human mutation and variation. *The American Journal of Human Genetics*, 91(6):1033–1040.

Kreisel, K., Engqvist, M. K. M., Kalm, J., Thompson, L. J., Boström, M., Navarrete, C., McDonald, J. P., Larsson, E., Woodgate, R., and Clausen, A. R. (2019). Dna polymerase ηcontributes to genome-wide lagging strand synthesis. *Nucleic Acids Res*, 47(5):2425–2435.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A., Wang, X., ClaussnitzerYaping Liu, M., Coarfa, C., Alan Harris, R., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., David Hawkins, R., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Scott Hansen, R., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Abdennur, N., Adli, M., Akerman, M., Barrera, L., Antosiewicz-Bourget, J., Ballinger, T., Barnes, M. J., Bates, D., Bell, R. J. A., Bennett, D. A., Bianco, K., Bock, C., Boyle, P., Brinchmann, J., Caballero-Campo, P., Camahort, R., Carrasco-Alfonso, M. J., Charnecki, T., Chen, H., Chen, Z., Cheng, J. B., Cho, S., Chu, A., Chung, W.-Y., Cowan, C., Athena Deng, Q., Deshpande, V., Diegel, M., Ding, B., Durham, T., Echipare, L., Edsall, L., Flowers, D., Genbacev-Krtolica, O., Gifford, C., Gillespie, S., Giste, E., Glass, I. A., Gnirke, A., Gormley, M., Gu, H., Gu, J., Hafler, D. A., Hangauer, M. J., Hariharan, M., Hatan, M., Haugen, E., He, Y., Heimfeld, S., Herlofsen, S., Hou, Z., Humbert, R., Issner, R., Jackson, A. R., Jia, H., Jiang, P., Johnson, A. K., Kadlecek, T., Kamoh, B., Kapidzic, M., Kent, J., Kim, A., Kleinewietfeld, M., Klugman, S., Krishnan, J., Kuan, S., Kutyavin, T., Lee, A.-Y., Lee, K., Li, J., Li, N., Li, Y., Ligon, K. L., Lin, S., Lin, Y., Liu, J., Liu, Y., Luckey, C. J., Ma, Y. P., Maire, C., Marson, A., Mattick, J. S., Mayo, M., McMaster, M., Metsky, H., Mikkelsen, T., Miller, D., Miri, M., Mukame, E., Nagarajan, R. P., Neri, F., Nery, J.,

Nguyen, T., O'Geen, H., Paithankar, S., Papayannopoulou, T., Pelizzola, M., Plettner, P., Propson, N. E., Raghuraman, S., Raney, B. J., Raubitschek, A., Reynolds, A. P., Richards, H., Riehle, K., Rinaudo, P., Robinson, J. F., Rockweiler, N. B., Rosen, E., Rynes, E., Schein, J., Sears, R., Sejnowski, T., Shafer, A., Shen, L., Shoemaker, R., Sigaroudinia, M., Slukvin, I., Stehling-Sun, S., Stewart, R., Subramanian, S. L., Suknuntha, K., Swanson, S., Tian, S., Tilden, H., Tsai, L., Urich, M., Vaughn, I., Vierstra, J., Vong, S., Wagner, U., Wang, H., Wang, T., Wang, Y., Weiss, A., Whitton, H., Wildberg, A., Witt, H., Won, K.-J., Xie, M., Xing, X., Xu, I., Xuan, Z., Ye, Z., Yen, C.-a., Yu, P., Zhang, X., Zhang, X., Zhao, J., Zhou, Y., Zhu, J., Zhu, Y., Ziegler, S., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., Kellis, M., Ziller, M. J., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sandstrom, R. S., Eaton, M. L., Pfenning, A. R., Claussnitzer, M., Liu, Y., Harris, R. A., Epstein, C. B., Hawkins, R. D., Mungall, A. J., Canfield, T. K., Hansen, R. S., Sabo, P. J., Bansal, M. S., Consortium, R. E., analysis coordination, I., analysis leads (equal contributors), I., production, D., processing leads (equal contributors), analysis co leads, I., Analysis, production contributors, principal investigators, C., program management, S., and investigators, P. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Kunkel, T. A. (2009). Evolving views of dna replication (in)fidelity. *Cold Spring Harbor symposia on quantitative biology*, 74:91–101.

Kunkel, T. A. and Bebenek, K. (2000). Dna replication fidelity. *Annual review of biochemistry*, 69(1):497–529.

Lal, A., Liu, K., Tibshirani, R., Sidow, A., and Ramazzotti, D. (2020). De novo mutational signature discovery in tumor genomes using sparsesignatures. *bioRxiv*.

Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–197.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., and et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

Lee, D. D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 535–541, Cambridge, MA, USA. MIT Press.

Lee, J., Lee, A. J., Lee, J.-K., Park, J., Kwon, Y., Park, S., Chun, H., Ju, Y. S., and Hong, D. (2018). Mutalisk: a web-based somatic mutation analyis toolkit for genomic, transcriptional and epigenomic signatures. *Nucleic Acids Research*, 46(W1):W102–W108.

Lee, J., Thompson, J. R., Botuyan, M. V., and Mer, G. (2008). Distinct binding modes specify the recognition of methylated histones h3k4 and h4k20 by jmjd2a-tudor. *Nature structural & molecular biology*, 15(1):109–111.

Lee, K. K. and Workman, J. L. (2007). Histone acetyltransferase complexes: one size doesn't fit all. *Nature Reviews Molecular Cell Biology*, 8(4):284–295.

Lee, Y.-C., Cai, Y., Mu, H., Broyde, S., Amin, S., Chen, X., Min, J.-H., and Geacintov, N. E. (2014). The relationships between xpc binding to conformationally diverse dna adducts and their excision by the human ner system: is there a correlation? *DNA repair*, 19:55–63.

Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R. J., Sanders, M. A., Moore, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., Robinson, P., Coorens, T. H. H., O'Neill, L., Alder, C., Wang, J., Fitzgerald, R. C., Zilbauer, M., Coleman, N., Saeb-Parsy, K., Martincorena, I., Campbell, P. J., and Stratton, M. R. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. *Nature*, 574(7779):532–537.

Lehmann, A. R. (2011). Dna polymerases and repair synthesis in ner in human cells. *DNA Repair (Amst)*, 10(7):730–733.

Letouzé, E., Shinde, J., Renault, V., Couchy, G., Blanc, J.-F., Tubacher, E., Bayard, Q., Bacq, D., Meyer, V., Semhoun, J., et al. (2017). Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nature communications*, 8(1):1315.

Li, G., Levitus, M., Bustamante, C., and Widom, J. (2005). Rapid spontaneous accessibility of nucleosomal dna. *Nat Struct Mol Biol*, 12(1):46–53.

Li, J., Harris, R. A., Cheung, S. W., Coarfa, C., Jeong, M., Goodell, M. A., White, L. D., Patel, A., Kang, S.-H., Shaw, C., Chinault, A. C., Gambin, T., Gambin, A., Lupski, J. R., and Milosavljevic, A. (2012). Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. *PLoS genetics*, 8(5):e1002692–e1002692.

Li, Y., Roberts, N. D., Wala, J. A., Shapira, O., Schumacher, S. E., Kumar, K., Khurana, E., Waszak, S., Korbel, J. O., Haber, J. E., Imielinski, M., Akdemir, K. C., Alvarez, E. G., Baez-Ortega, A., Beroukhim, R., Boutros, P. C., Bowtell, D. D. L., Brors, B., Burns, K. H., Campbell, P. J., Chan, K., Chen, K., Cortés-Ciriano, I., Dueso-Barroso, A., Dunford, A. J., Edwards, P. A., Estivill, X., Etemadmoghadam, D., Feuerbach, L., Fink, J. L., Frenkel-Morgenstern, M., Garsed, D. W., Gerstein, M., Gordenin, D. A., Haan, D., Haber, J. E., Hess, J. M., Hutter, B., Jones, D. T. W., Ju, Y. S., Kazanov, M. D., Klimczak, L. J., Koh, Y., Korbel, J. O., Lee, E. A., Lee, J. J.-K., Lynch, A. G., Macintyre, G., Markowetz, F., Martincorena, I., Martinez-Fundichely, A., Meyerson, M., Miyano, S., Nakagawa, H., Navarro, F. C. P., Ossowski, S., Park, P. J., Pearson, J. V., Puiggròs, M., Rippe, K., Roberts, N. D., Roberts, S. A., Rodriguez-Martin, B., Schumacher, S. E., Scully, R., Shackleton, M., Sidiropoulos, N., Sieverling, L., Stewart, C., Torrents, D., Tubio, J. M. C., Villasante, I., Waddell, N., Wala, J. A., Weischenfeldt, J., Yang, L., Yao, X., Yoon, S.-S., Zamora, J., Zhang, C.-Z., Campbell, P. J., Group, P. S. V. W., and Consortium, P. (2020). Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121.

Li, Y. F., Kim, S.-T., and Sancar, A. (1993). Evidence for lack of dna photoreactivating enzyme in humans. *Proceedings of the National Academy of Sciences*, 90(10):4389–4393.

Lieber, M. R. (2008). The mechanism of human nonhomologous dna end joining. *Journal of Biological Chemistry*, 283(1):1–5.

Lindahl, T. et al. (1993). Instability and decay of the primary structure of dna. *nature*, 362(6422):709–715.

Lindahl, T. and Nyberg, B. (1974). Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry*, 13(16):3405–3410.

Little, J. B. (1993). Cellular, molecular, and carcinogenic effects of radiation. *Hematology/Oncology Clinics*, 7(2):337–352.

Lomax, M., Folkes, L., and O'Neill, P. (2013). Biological consequences of radiation-induced dna damage: relevance to radiotherapy. *Clinical oncology*, 25(10):578–585.

Lovett, S. T. (2007). Polymerase switching in dna replication. *Molecular Cell*, 27(4):523–526.

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 åresolution. *Nature*, 389(6648):251–260.

Martincorena, I., Fowler, J. C., Wabik, A., Lawson, A. R., Abascal, F., Hall, M. W., Cagan, A., Murai, K., Mahbubani, K., Stratton, M. R., et al. (2018). Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917.

McCulloch, S. D., Kokoska, R. J., Masutani, C., Iwai, S., Hanaoka, F., and Kunkel, T. A. (2004). Preferential cis–syn thymine dimer bypass by dna polymerase ηoccurs with biased fidelity. *Nature*, 428(6978):97–100.

McCulloch, S. D. and Kunkel, T. A. (2008). The fidelity of dna synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res*, 18(1):148–161.

Meier, B., Volkova, N. V., Hong, Y., Schofield, P., Campbell, P. J., Gerstung, M., and Gartner, A. (2018). Mutational signatures of DNA mismatch repair deficiency in c. elegans and human cancers. *Genome Res.*, 28(5):666–675.

Milazzo, G., Mercatelli, D., Di Muzio, G., Triboli, L., De Rosa, P., Perini, G., and Giorgi, F. M. (2020). Histone deacetylases (hdacs): Evolution, specificity, role in transcriptional complexes, and pharmacological actionability. *Genes (Basel)*, 11(5).

Mimitou, E. P. and Symington, L. S. (2009). Nucleases and helicases take center stage in homologous recombination. *Trends in Biochemical Sciences*, 34(5):264–272.

Min, J., Zhang, Y., and Xu, R.-M. (2003). Structural basis for specific binding of polycomb chromodomain to histone h3 methylated at lys 27. *Genes Dev*, 17(15):1823–1828.

Mišković, K., Bujak, M., Baus Lončar, M., and Glavaš-Obrovac, L. (2013). Antineoplastic dna-binding compounds: intercalating and minor groove binding drugs. *Arh Hig Rada Toksikol*, 64(4):593–602.

Miura, K., Kinouchi, M., Ishida, K., Fujibuchi, W., Naitoh, T., Ogawa, H., Ando, T., Yazaki, N., Watanabe, K., Haneda, S., Shibata, C., and Sasaki, I. (2010). 5-fu metabolism in cancer and orally-administrable 5-fu drugs. *Cancers*, 2(3):1717–1730.

Moore, L., Leongamornlert, D., Coorens, T. H. H., Sanders, M. A., Ellis, P., Dentro, S. C., Dawson, K. J., Butler, T., Rahbari, R., Mitchell, T. J., Maura, F., Nangalia, J., Tarpey, P. S., Brunner, S. F., Lee-Six, H., Hooks, Y., Moody, S., Mahbubani, K. T., Jimenez-Linan, M., Brosens, J. J., Iacobuzio-Donahue, C. A., Martincorena, I., Saeb-Parsy, K., Campbell,

P. J., and Stratton, M. R. (2020). The mutational landscape of normal human endometrial epithelium. *Nature*, 580(7805):640–646.

Morganella, S., Alexandrov, L. B., Glodzik, D., Zou, X., Davies, H., Staaf, J., Sieuwerts, A. M., Brinkman, A. B., Martin, S., Ramakrishna, M., and et al. (2016). The topography of mutational processes in breast cancer genomes. *Nature Communications*, 7(1).

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (aid), a potential rna editing enzyme. *Cell*, 102(5):553–563.

Narita, T., Yamaguchi, Y., Yano, K., Sugimoto, S., Chanarat, S., Wada, T., Kim, D.-k., Hasegawa, J., Omori, M., Inukai, N., Endoh, M., Yamada, T., and Handa, H. (2003). Human transcription elongation factor nelf: identification of novel subunits and reconstitution of the functionally active complex. *Mol Cell Biol*, 23(6):1863–1873.

Nick McElhinny, S. A., Snowden, C. M., McCarville, J., and Ramsden, D. A. (2000). Ku recruits the xrcc4-ligase iv complex to dna ends. *Mol Cell Biol*, 20(9):2996–3003.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., and et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjærde, O. C., Langerød, A., Ringnér, M., Ahn, S.-M., Boyault, S., Brock, J. E., Broeks, A., Butler, A., Desmedt, C., Dirix, L., Dronov, S., Fatima, A., Foekens, J. A., Gerstung, M., Hooijer, G. K. J., Jang, S. J., Jones, D. R., Kim, H.-Y., King, T. A., Krishnamurthy, S., Lee, H. J., Lee, J.-Y., Li, Y., McLaren, S., Menzies, A., Mustonen, V., O'Meara, S., Pauporté, I., Pivot, X., Purdie, C. A., Raine, K., Ramakrishnan, K., Rodríguez-González, F. G., Romieu, G., Sieuwerts, A. M., Simpson, P. T., Shepherd, R., Stebbings, L., Stefansson, O. A., Teague, J., Tommasi, S., Treilleux, I., Van den Eynden, G. G., Vermeulen, P., Vincent-Salomon, A., Yates, L., Caldas, C., van't Veer, L., Tutt, A., Knappskog, S., Tan, B. K. T., Jonkers, J., Borg, Å., Ueno, N. T., Sotiriou, C., Viari, A., Futreal, P. A., Campbell, P. J., Span, P. N., Van Laere, S., Lakhani, S. R., Eyfjord, J. E., Thompson, A. M., Birney, E., Stunnenberg, H. G., van de Vijver, M. J., Martens, J. W. M., Børresen-Dale, A.-L., Richardson, A. L., Kong, G., Thomas, G., and Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54.

Nik-Zainal, S., Wedge, D. C., Alexandrov, L. B., Petljak, M., Butler, A. P., Bolli, N., Davies, H. R., Knappskog, S., Martin, S., Papaemmanuil, E., Ramakrishna, M., Shlien, A., Simonic, I., Xue, Y., Tyler-Smith, C., Campbell, P. J., and Stratton, M. R. (2014). Association of a germline copy number polymorphism of apobec3a and apobec3b with burden of putative apobec-dependent mutations in breast cancer. *Nat Genet*, 46(5):487–491.

Nitiss, J. L. (2009). Targeting dna topoisomerase ii in cancer chemotherapy. *Nature Reviews Cancer*, 9(5):338.

Ogi, T., Limsirichaikul, S., Overmeer, R., Volker, M., Takenaka, K., Cloney, R., Nakazawa, Y., Niimi, A., Miki, Y., Jaspers, N. G., Mullenders, L. H. F., Yamashita, S., Fousteri, M. I., and Lehmann, A. R. (2010). Three dna polymerases, recruited by different mechanisms, carry out ner repair synthesis in human cells. *Mol Cell*, 37(5):714–727.

Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257.

Owen-Hughes, T. and Gkikopoulos, T. (2012). Making sense of transcribing chromatin. *Current opinion in cell biology*, 24(3):296–304.

Pâques, F. and Haber, J. E. (1999). Multiple pathways of recombination induced by double-strand breaks in saccharomyces cerevisiae. *Microbiol Mol Biol Rev*, 63(2):349–404.

Parkash, V., Kulkarni, Y., ter Beek, J., Shcherbakova, P. V., Kamerlin, S. C. L., and Johansson, E. (2019). Structural consequence of the most frequently recurring cancer-associated substitution in DNA polymerase . *Nature Communications*, 10(1).

Petljak, M., Alexandrov, L. B., Brammeld, J. S., Price, S., Wedge, D. C., Grossmann, S., Dawson, K. J., Ju, Y. S., Iorio, F., Tubio, J. M. C., Koh, C. C., Georgakopoulos-Soares, I., Rodríguez-Martín, B., Otlu, B., O'Meara, S., Butler, A. P., Menzies, A., Bhosle, S. G., Raine, K., Jones, D. R., Teague, J. W., Beal, K., Latimer, C., O'Neill, L., Zamora, J., Anderson, E., Patel, N., Maddison, M., Ng, B. L., Graham, J., Garnett, M. J., McDermott, U., Nik-Zainal, S., Campbell, P. J., and Stratton, M. R. (2019). Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell*, 176(6):1282–1294.e20.

Petruseva, I., Evdokimov, A., and Lavrik, O. (2014). Molecular mechanism of global genome nucleotide excision repair. *Acta Naturae ( )*, 6(1 (20)).

Pham, P., Bransteitter, R., Petruska, J., and Goodman, M. F. (2003). Processive aid-catalysed cytosine deamination on single-stranded dna simulates somatic hypermutation. *Nature*, 424(6944):103–107.

Pich, O., Muiños, F., Lolkema, M. P., Steeghs, N., Gonzalez-Perez, A., and Lopez-Bigas, N. (2019). The mutational footprints of cancer therapies. *Nature Genetics*, 51(12):1732–1740.

Pich, O., Muiños, F., Sabarinathan, R., Reyes-Salazar, I., Gonzalez-Perez, A., and Lopez-Bigas, N. (2018). Somatic and germline mutation periodicity follow the orientation of the dna minor groove around nucleosomes. *Cell*, 175(4):1074 – 1087.e18.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., et al. (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191.

Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M.-L., Beare, D., Lau, K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordoñez, G. R., Mudie, L. J., Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J. W., Mangion, J., Sun, Y. A.,

McLaughlin, S. F., Peckham, H. E., Tsung, E. F., Costa, G. L., Lee, C. C., Minna, J. D., Gazdar, A., Birney, E., Rhodes, M. D., McKernan, K. J., Stratton, M. R., Futreal, P. A., and Campbell, P. J. (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278):184–190.

Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., van Hoeck, A., Wood, H. M., Nomburg, J., Gurjao, C., Manders, F., Dalmasso, G., Stege, P. B., Paganelli, F. L., Geurts, M. H., Beumer, J., Mizutani, T., Miao, Y., van der Linden, R., van der Elst, S., Ambrose, J. C., Arumugam, P., Baple, E. L., Bleda, M., Boardman-Pretty, F., Boissiere, J. M., Boustred, C. R., Brittain, H., Caulfield, M. J., Chan, G. C., Craig, C. E. H., Daugherty, L. C., de Burca, A., Devereau, A., Elgar, G., Foulger, R. E., Fowler, T., Furió-Tarí, P., Hackett, J. M., Halai, D., Hamblin, A., Henderson, S., Holman, J. E., Hubbard, T. J. P., Ibáñez, K., Jackson, R., Jones, L. J., Kasperaviciute, D., Kayikci, M., Lahnstein, L., Lawson, L., Leigh, S. E. A., Leong, I. U. S., Lopez, F. J., Maleady-Crowe, F., Mason, J., McDonagh, E. M., Moutsianas, L., Mueller, M., Murugaesu, N., Need, A. C., Odhams, C. A., Patch, C., Perez-Gil, D., Polychronopoulos, D., Pullinger, J., Rahim, T., Rendon, A., Riesgo-Ferreiro, P., Rogers, T., Ryten, M., Savage, K., Sawant, K., Scott, R. H., Siddiq, A., Sieghart, A., Smedley, D., Smith, K. R., Sosinsky, A., Spooner, W., Stevens, H. E., Stuckey, A., Sultana, R., Thomas, E. R. A., Thompson, S. R., Tregidgo, C., Tucci, A., Walsh, E., Watters, S. A., Welland, M. J., Williams, E., Witkowska, K., Wood, S. M., Zarowiecki, M., Garcia, K. C., Top, J., Willems, R. J. L., Giannakis, M., Bonnet, R., Quirke, P., Meyerson, M., Cuppen, E., van Boxtel, R., Clevers, H., and Consortium, G. E. R. (2020). Mutational signature in colorectal cancer caused by genotoxic pks+ e. coli. *Nature*, 580(7802):269–273.

Pluciennik, A., Dzantiev, L., Iyer, R. R., Constantin, N., Kadyrov, F. A., and Modrich, P. (2010). Pcna function in the activation and strand direction of mutlα endonuclease in mismatch repair. *Proceedings of the National Academy of Sciences*, 107(37):16066–16071.

Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M. S., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J. A., and Sunyaev, S. R. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518(7539):360–364.

Polak, P., Kim, J., Braunstein, L. Z., Karlic, R., Haradhavala, N. J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K. W., et al. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature genetics*, 49(10):1476.

Polak, P., Lawrence, M. S., Haugen, E., Stoletzki, N., Stojanov, P., Thurman, R. E., Garraway, L. A., Mirkin, S., Getz, G., Stamatoyannopoulos, J. A., and Sunyaev, S. R. (2014). Reduced local mutation density in regulatory dna of cancer genomes is linked to dna repair. *Nature biotechnology*, 32(1):71–75.

Poon, S. L., McPherson, J. R., Tan, P., Teh, B. T., and Rozen, S. G. (2014). Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. *Genome Med.*, 6(3):24.

Priestley, P., Baber, J., Lolkema, M., Steeghs, N., de Bruijn, E., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., Roepman, P., et al. (2018). Pan-cancer whole genome analyses of metastatic solid tumors. *bioRxiv*, page 415133.

Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L., and Sweet, R. M. (1993). Crystal structure of globular domain of histone h5 and its implications for nucleosome binding. *Nature*, 362(6417):219–223.

Ransom, M., Dennehey, B. K., and Tyler, J. K. (2010). Chaperoning histones during dna replication and repair. *Cell*, 140(2):183–195.

Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G. V., Carter, S. L., Saksena, G., Harris, S., Shah, R. R., Resnick, M. A., Getz, G., and Gordenin, D. A. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.*, 45(9):970–976.

Robinson, P. J. J., Fairall, L., Huynh, V. A. T., and Rhodes, D. (2006). Em measurements define the dimensions of the "30-nm" chromatin fiber: Evidence for a compact, interdigitated structure. *Proceedings of the National Academy of Sciences*, 103(17):6506–6511.

Robinson, P. S., Coorens, T. H. H., Palles, C., Mitchell, E., Abascal, F., Olafsson, S., Lee, B., Lawson, A. R. J., Lee-Six, H., Moore, L., Sanders, M. A., Hewinson, J., Martin, L., Pinna, C. M. A., Galvotti, S., Campbell, P. J., Martincorena, I., Tomlinson, I., and Stratton, M. R. (2020). Elevated somatic mutation burdens in normal human cells due to defective dna polymerases. *bioRxiv*, page 2020.06.23.167668.

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.*, 17:31.

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to dna. *Nature*, 532(7598):264–267.

Saha, A., Wittmeyer, J., and Cairns, B. R. (2006). Chromatin remodelling: the industrial revolution of dna around histones. *Nature Reviews Molecular Cell Biology*, 7(6):437–447.

Sankar, T. S., Wastuwidyaningtyas, B. D., Dong, Y., Lewis, S. A., and Wang, J. D. (2016). The nature of mutations induced by replication–transcription collisions. *Nature*, 535(7610):178.

Schalch, T., Duda, S., Sargent, D. F., and Richmond, T. J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, 436(7047):138–141.

Schärer, O. D. (2013). Nucleotide excision repair in eukaryotes. *Cold Spring Harbor perspectives in biology*, 5(10):a012609–a012609.

Schuster-Böckler, B. and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488:504 EP –.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Scully, R., Panday, A., Elango, R., and Willis, N. A. (2019). Dna double-strand break repair-pathway choice in somatic mammalian cells. *Nature Reviews Molecular Cell Biology*, 20(11):698–714.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J.-P. Z., and Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778.

Serra-Cardona, A. and Zhang, Z. (2018). Replication-coupled nucleosome assembly in the passage of epigenetic information and cell identity. *Trends in biochemical sciences*, 43(2):136–148.

Shen, J.-C., Rideout, William M., I., and Jones, P. A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded dna. *Nucleic Acids Research*, 22(6):972–976.

Shinbrot, E., Henninger, E. E., Weinhold, N., Covington, K. R., Göksenin, A. Y., Schultz, N., Chao, H., Doddapaneni, H., Muzny, D. M., Gibbs, R. A., Sander, C., Pursell, Z. F., and Wheeler, D. A. (2014). Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res.*, 24(11):1740–1750.

Smith, K. S., Liu, L. L., Ganesan, S., Michor, F., and De, S. (2017). Nuclear topology modulates the mutational landscapes of cancer genomes. *Nature Structural & Molecular Biology*, 24(11):1000–1006.

Spangler, L., Wang, X., Conaway, J. W., Conaway, R. C., and Dvir, A. (2001). Tfiih action in transcription initiation and promoter escape requires distinct regions of downstream promoter dna. *Proceedings of the National Academy of Sciences*, 98(10):5544.

Spencer, J. P., Jenner, A., Aruoma, O. I., Cross, C. E., Wu, R., and Halliwell, B. (1996). Oxidative dna damage in human respiratory tract epithelial cells. time course in relation to dna strand breakage. *Biochem Biophys Res Commun*, 224(1):17–22.

Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., and Sunyaev, S. R. (2009). Human mutation rate associated with dna replication timing. *Nat Genet*, 41(4):393–395.

Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765):41–45.

Supek, F. and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*, 521(7550):81–84.

Supek, F. and Lehner, B. (2017). Clustered mutation signatures reveal that error-prone dna repair targets mutations to active genes. *Cell*, 170(3):534–547.

Swan, M. K., Johnson, R. E., Prakash, L., Prakash, S., and Aggarwal, A. K. (2009). Structural basis of high-fidelity dna synthesis by yeast dna polymerase delta. *Nat Struct Mol Biol*, 16(9):979–986.

Takata, K.-i., Shimizu, T., Iwai, S., and Wood, R. D. (2006). Human dna polymerase n (poln) is a low fidelity enzyme capable of error-free bypass of 5s-thymine glycol. *J Biol Chem*, 281(33):23445–23455.

Taylor, B. J., Nik-Zainal, S., Wu, Y. L., Stebbings, L. A., Raine, K., Campbell, P. J., Rada, C., Stratton, M. R., and Neuberger, M. S. (2013). DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife*, 2:e00534.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.

Tomkova, M., Tomek, J., Kriaucionis, S., and Schuster-Böckler, B. (2018). Mutational signature distribution varies with dna replication timing and strand asymmetry. *Genome biology*, 19(1):129.

Torres, C. M., Biran, A., Burney, M. J., Patel, H., Henser-Brownhill, T., Cohen, A.-H. S., Li, Y., Ben-Hamo, R., Nye, E., Spencer-Dene, B., Chakravarty, P., Efroni, S., Matthews, N., Misteli, T., Meshorer, E., and Scaffidi, P. (2016). The linker histone h1.0 generates epigenetic and functional intratumor heterogeneity. *Science*, 353(6307):aaf1644.

Toyota, M. and Suzuki, H. (2010). Epigenetic drivers of genetic alterations. *Adv Genet*, 70:309–323.

Tsesmetzis, N., Paulin, C. B. J., Rudd, S. G., and Herold, N. (2018). Nucleobase and nucleoside analogues: Resistance and re-sensitisation at the level of pharmacokinetics, pharmacodynamics and metabolism. *Cancers*, 10(7):240.

van Steensel, B. and Furlong, E. E. M. (2019). The role of transcription in shaping the spatial organization of the genome. *Nature Reviews Molecular Cell Biology*, 20(6):327–337.

Venkatesh, S. and Workman, J. L. (2013). Set2 mediated h3 lysine 36 methylation: regulation of transcription elongation and implications in organismal development. *Wiley interdisciplinary reviews. Developmental biology*, 2(5):685–700.

Viel, A., Bruselles, A., Meccia, E., Fornasarig, M., Quaia, M., Canzonieri, V., Policicchio, E., Urso, E. D., Agostini, M., Genuardi, M., et al. (2017). A specific mutational signature associated with dna 8-oxoguanine persistence in mutyh-defective colorectal cancer. *EBioMedicine*, 20:39–49.

Volkova, N. V., Meier, B., González-Huici, V., Bertolini, S., Gonzalez, S., Abascal, F., Martincorena, I., Campbell, P. J., Gartner, A., and Gerstung, M. (2019). Mutational signatures are jointly shaped by dna damage and repair. *bioRxiv*, page 686295.

Wakelin, L. P. (1986). Polyfunctional dna intercalating agents. *Medicinal research reviews*, 6(3):275–340.

Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M. U., Ohgi, K. A., Glass, C. K., Rosenfeld, M. G., and Fu, X.-D. (2011a). Reprogramming transcription by distinct classes of enhancers functionally defined by erna. *Nature*, 474(7351):390–394.

Wang, J. C. (2002). Cellular roles of dna topoisomerases: a molecular perspective. *Nature Reviews Molecular Cell Biology*, 3(6):430–440.

Wang, W., Hellinga, H. W., and Beese, L. S. (2011b). Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. *Proceedings of the National Academy of Sciences*, 108(43):17644–17648.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Weinberg, R. A. (2006). *The Biology of Cancer*. Garland Science.

Weinberg, R. A. (2013). *The Biology of Cancer*. W.W. Norton.

West, A. G., Gaszner, M., and Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. *Genes Dev*, 16(3):271–288.

Wolfe, K. H., Sharp, P. M., and Li, W. H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–285.

Xie, A., Kwok, A., and Scully, R. (2009). Role of mammalian mre11 in classical and alternative nonhomologous end joining. *Nature structural & molecular biology*, 16(8):814–818.

Xing, X., Kane, D. P., Bulock, C. R., Moore, E. A., Sharma, S., Chabes, A., and Shcherbakova, P. V. (2019). A recurrent cancer-associated substitution in DNA polymerase produces a hyperactive enzyme. *Nature Communications*, 10(1).

Xue, W. and Warshawsky, D. (2005). Metabolic activation of polycyclic and heterocyclic aromatic hydrocarbons and dna damage: a review. *Toxicology and applied pharmacology*, 206(1):73–93.

Yang, H., Li, Q., Fan, J., Holloman, W. K., and Pavletich, N. P. (2005). The brca2 homologue brh2 nucleates rad51 filament formation at a dsdna–ssdna junction. *Nature*, 433(7026):653–657.

Yarosh, D. B. (1985). The role of o6-methylguanine-dna methyltransferase in cell survival, mutagenesis and carcinogenesis. *Mutation Research/DNA Repair Reports*, 145(1):1–16.

Yekezare, M., Gómez-González, B., and Diffley, J. F. X. (2013). Controlling dna replication origins in response to dna damage – inhibit globally, activate locally. *Journal of Cell Science*, 126(6):1297–1306.

Yoshida, K., Gowers, K. H. C., Lee-Six, H., Chandrasekharan, D. P., Coorens, T., Maughan, E. F., Beal, K., Menzies, A., Millar, F. R., Anderson, E., Clarke, S. E., Pennycuick, A., Thakrar, R. M., Butler, C. R., Kakiuchi, N., Hirano, T., Hynds, R. E., Stratton, M. R., Martincorena, I., Janes, S. M., and Campbell, P. J. (2020). Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature*, 578(7794):266–272.

Zheng, C. L., Wang, N. J., Chung, J., Moslehi, H., Sanborn, J. Z., Hur, J. S., Collisson, E. A., Vemula, S. S., Naujokas, A., Chiotti, K. E., Cheng, J. B., Fassihi, H., Blumberg, A. J., Bailey, C. V., Fudem, G. M., Mihm, F. G., Cunningham, B. B., Neuhaus, I. M., Liao, W., Oh, D. H., Cleaver, J. E., LeBoit, P. E., Costello, J. F., Lehmann, A. R., Gray, J. W., Spellman, P. T., Arron, S. T., Huh, N., Purdom, E., and Cho, R. J. (2014). Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. *Cell Rep.*, 9(4):1228–1234.

Zou, X., Owusu, M., Harris, R., Jackson, S. P., Loizou, J. I., and Nik-Zainal, S. (2018). Validating the concept of mutational signatures with isogenic cell models. *Nature communications*, 9(1):1744.